Analyse numérique

Catherine Bolley

Catherine Bolley

Analyse numérique

Centrale Nantes

Table des matières

1

Généralités sur les matrices

Notations et rappels

Dans ce document, l'ensemble $\mathbb{K}$ est un corps commutatif, $\mathbb{R}$ ou $\mathbb{C}$. $\mathbb{K}^n$ est l'espace vectoriel des vecteurs (colonnes) à n lignes à coefficients dans $\mathbb{K}$. Si $\mathbf{u} \in \mathbb{K}^n$, on notera $\mathbf{u} = (u_i)_{i=1,\dots,n}$ où les u_i sont les coefficients de $\mathbf{u}$. $\mathbb{K}^{n,m}$ est l'espace vectoriel des matrices à n lignes et m colonnes à coefficients dans $\mathbb{K}$. Si $\mathbf{A} \in \mathbb{K}^{n,m}$ a pour coefficients a_{ij}, on notera :

$$\mathbf{A} = (a_{ij})_{i,j}, \quad i = 1,\dots,n, \quad j = 1,\dots,m \tag{1.1}$$

où i désigne l'indice de ligne et j, l'indice de colonne. La matrice $\mathbf{A}^\top \in \mathbb{K}^{m,n}$ désigne la transposée de la matrice $\mathbf{A}$ et ses coefficients vérifient :

$$a_{ij}^\top = a_{ji}, \quad i = 1,\dots,n, \quad j = 1,\dots,m \tag{1.2}$$

La matrice $\mathbf{A}^* \in \mathbb{K}^{n,m}$ est l'adjointe de la matrice $\mathbf{A}$:

$$\mathbf{A}^* = \bar{\mathbf{A}}^\top \tag{1.3}$$

Matrices carrées On s'intéresse au cas $m = n$ et $\mathbf{A} \in \mathbb{C}^{n,n}$. On note λ_i, $i = 1,\dots,n$ les n valeurs propres dans $\mathbb{C}$ de $\mathbf{A}$.

▷ **Définition 1.1 — Rayon spectral.** On appelle rayon spectral de $\mathbf{A}$, le réel noté $\rho(\mathbf{A})$ défini par :

$$\rho(\mathbf{A}) = \max_{i=1,\dots,n} |\lambda_i| \tag{1.4}$$

▷ **Définition 1.2 — Trace.** La trace de la matrice $\mathbf{A}$ est donnée par :

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} \tag{1.5}$$

On a en particulier :

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i \tag{1.6}$$

▷ **Définition 1.3 — Matrice diagonale.** Une matrice carrée $\mathbf{A} \in \mathbb{K}^{n,n}$ est dite diagonale lorsque $a_{ij} = 0$ pour $i \neq j$.

▷ **Définition 1.4 — Matrice bande.** Une matrice carrée $\mathbf{A} \in \mathbb{K}^{n,n}$ est dite bande (p, q) lorsque $a_{ij} = 0$ pour $i \geqslant j + p$ et pour $j \geqslant i + q$. Elle est alors de la forme suivante :

$$\mathbf{A} = \begin{array}{c} \overbrace{}^{q} \\ \left.\begin{array}{c} p \\ \\ \\ \\ \end{array}\right\{ \end{array} \begin{bmatrix} \bullet & \bullet & \bullet & 0 & \cdots & 0 \\ \bullet & \bullet & \bullet & \bullet & \ddots & \vdots \\ 0 & \bullet & \bullet & \bullet & \bullet & 0 \\ \vdots & \ddots & \bullet & \bullet & \bullet & \bullet \\ 0 & \cdots & 0 & \bullet & \bullet & \bullet \end{bmatrix} \tag{1.7}$$

▷ **Définition 1.5** Une matrice carrée $\mathbf{A} \in \mathbb{K}^{n,n}$ est dite $2\ell + 1$ diagonale avec $\ell \in \mathbb{N}^*$ si c'est une matrice bande $(\ell + 1, \ell + 1)$, c'est-à-dire si $a_{ij} = 0$ pour $i \geqslant j + \ell + 1$ et pour $j \geqslant i + \ell + 1$.

Si $\mathbf{u} \in \mathbb{K}^n$ et $\mathbf{v} \in \mathbb{K}^n$, la quantité $(\mathbf{u}, \mathbf{v})$ désigne :

— le produit scalaire euclidien si $\mathbb{K} = \mathbb{R}$:

$$(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n} u_i v_i \tag{1.8}$$

— le produit scalaire hermitien si $\mathbb{K} = \mathbb{C}$:

$$(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n} u_i \bar{v}_i \tag{1.9}$$

▷ **Définition 1.6 — Matrice hermitienne.** Une matrice carrée $\mathbf{A} \in \mathbb{K}^{n,n}$ est dite hermitienne lorsque :

$$\mathbf{A}^* = \mathbf{A} \tag{1.10}$$

Quand $\mathbb{K} = \mathbb{R}$, elle est dite symétrique.

▷ **Définition 1.7 — Matrice positive.** Une matrice hermitienne $\mathbf{A} \in \mathbb{K}^{n,n}$ est dite positive lorsque :

$$\forall \mathbf{x} \in \mathbb{K}^n, \quad (\mathbf{A}\mathbf{x}, \mathbf{x}) \geqslant 0 \tag{1.11}$$

▷ **Définition 1.8 — Matrice définie positive.** Une matrice hermitienne $\mathbf{A} \in \mathbb{K}^{n,n}$ est dite définie positive lorsque :

$$\forall \mathbf{x} \in \mathbb{K}^n, \quad \mathbf{x} \neq \mathbf{0}, \quad (\mathbf{A}\mathbf{x}, \mathbf{x}) > 0 \tag{1.12}$$

Théorème 1.9

1. Une matrice hermitienne $\mathbf{A}$ a toutes ses valeurs propres réelles et il existe une base de $\mathbb{C}^n$ de vecteurs propres de $\mathbf{A}$; la matrice $\mathbf{A}$ est donc diagonalisable. En particulier, une matrice réelle symétrique a toutes ses valeurs propres réelles.
2. Une matrice hermitienne est définie positive (resp. positive) si et seulement si toutes ses valeurs propres sont strictement positives (resp. positives ou nulles).
3. Pour toute matrice $\mathbf{A} \in \mathbb{C}^{n,n}$, il existe une matrice inversible $\mathbf{P} \in \mathbb{C}^{n,n}$ telle que la matrice $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ soit diagonale par bloc, chaque bloc étant une sous-matrice de Jordan $\mathbf{J}_p$

d'ordre n_p avec :

$$\mathbf{J}_p = \begin{bmatrix} \lambda_p & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_p \end{bmatrix} \tag{1.13}$$

▷ **Définition 1.10 — Matrice à diagonale strictement dominante.** Une matrice $\mathbf{A} \in \mathbb{R}^{n,n}$ est dite à diagonale strictement dominante si :

$$\forall i = 1, \ldots, n, \quad |a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \tag{1.14}$$

▷ **Proposition 1.11** Une matrice à diagonale strictement dominante est inversible.

Démonstration. Voir l'exercice 39.

1.1 Normes matricielles

▷ **Définition 1.12 — Norme matricielle.** On appelle norme matricielle sur $\mathbb{K}^{n,n}$, toute application de $\mathbb{K}^{n,n}$ dans $\mathbb{R}^+$, notée $\mathbf{A} \in \mathbb{K}^{n,n} \to ||\mathbf{A}||$, possédant les propriétés suivantes :

1. $||\mathbf{A}|| = 0 \quad \Leftrightarrow \quad \mathbf{A} = \mathbf{0}$
2. $\forall \lambda \in \mathbb{K}, \forall \mathbf{A} \in \mathbb{K}^{n,n}, ||\lambda \mathbf{A}|| = |\lambda| \, ||\mathbf{A}||$
3. $\forall \mathbf{A} \in \mathbb{K}^{n,n}, \forall \mathbf{B} \in \mathbb{K}^{n,n}, ||\mathbf{A} + \mathbf{B}|| \leqslant ||\mathbf{A}|| + ||\mathbf{B}||$
4. $\forall \mathbf{A} \in \mathbb{K}^{n,n}, \forall \mathbf{B} \in \mathbb{K}^{n,n}, ||\mathbf{A} \cdot \mathbf{B}|| \leqslant ||\mathbf{A}|| \cdot ||\mathbf{B}||$

■ **Exemple — Norme de Schur.** Soit $\mathbf{A} \in \mathbb{K}^{n,n}$; sa norme de Schur est définie par :

$$||\mathbf{A}||_S = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2} \tag{1.15}$$

▷ **Proposition 1.13** Étant donnée $||\cdot||$ une norme vectorielle sur $\mathbb{K}^n$, l'application encore notée $||\cdot||$ et définie par :

$$\mathbf{A} \in \mathbb{K}^{n,n} \to ||\mathbf{A}|| = \sup_{\mathbf{x} \in \mathbb{K}^n, \mathbf{x} \neq \mathbf{0}} \frac{||\mathbf{A}\mathbf{x}||}{||\mathbf{x}||} \tag{1.16}$$

est une norme matricielle. Cette norme vérifie :

$$||\mathbf{A}|| = \sup_{0 < ||\mathbf{x}|| \leqslant 1} ||\mathbf{A}\mathbf{x}|| = \sup_{||\mathbf{x}|| = 1} ||\mathbf{A}\mathbf{x}|| \tag{1.17}$$

▷ **Définition 1.14** La norme matricielle définie par (1.16) ou (1.17) est dite subordonnée à la norme vectorielle $||\cdot||$ donnée.

La proposition 1.13 résulte d'un théorème classique concernant les normes dans $\mathcal{L}(E, F)$ où E et F sont des espaces vectoriels sur le même corps $\mathbb{K}$ tels que $\dim E = m$ et $\dim F = n$.

Notations Pour p tel que $1 \leqslant p \leqslant +\infty$, on notera $||\mathbf{A}||_p$ la norme matricielle subordonnée à la norme vectorielle $||\mathbf{x}||_p$ définie par :

$$\mathbf{x} = (x_i)_{1\leqslant i\leqslant n} \rightarrow \begin{cases} ||\mathbf{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} & \text{si} \quad 1 \leqslant p < +\infty \\ ||\mathbf{x}||_\infty = \max_{i=1,\ldots,n} |x_i| & \text{si} \quad p = +\infty \end{cases} \tag{1.18}$$

▷ **Proposition 1.15** Si $\mathbf{A} = (a_{ij}) \in \mathbb{K}^{n,n}$, alors :

$$||\mathbf{A}||_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \tag{1.19}$$

$$||\mathbf{A}||_2 = \left(\rho(\mathbf{A}^*\mathbf{A})\right)^{1/2} \tag{1.20}$$

$$||\mathbf{A}||_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \tag{1.21}$$

Démonstration. Voir l'exercice 9.

Remarques

1. Une norme matricielle n'est pas nécessairement subordonnée à une norme vectorielle, comme par exemple la norme de Schur pour laquelle $||\mathbf{I}||_S = \sqrt{n}$;
2. Sous les hypothèses de la proposition 1.13, nous avons :
$$\forall \mathbf{x} \in \mathbb{K}^n, \quad \forall \mathbf{A} \in \mathbb{K}^{n,n}, \quad ||\mathbf{Ax}|| \leqslant ||\mathbf{A}|| \cdot ||\mathbf{x}|| \tag{1.22}$$
Inversement, étant donnée une norme matricielle $||\cdot||$ sur $\mathbb{K}^{n,n}$, il existe toujours une norme vectorielle $||\cdot||$ telle que (1.22) soit vérifiée. On peut choisir par exemple :
$$\forall \mathbf{x} \in \mathbb{K}^n, \quad ||\mathbf{x}|| = ||\begin{bmatrix} x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ x_n & 0 & \cdots & 0 \end{bmatrix}|| \tag{1.23}$$
L'axiome (4) de la définition 1.12 implique en effet que $||\mathbf{Ax}|| \leqslant ||\mathbf{A}||\,||\mathbf{x}||$.

Théorème 1.16 Soit $\mathbf{A} \in \mathbb{K}^{n,n}$ alors pour toute norme matricielle $||\cdot||$, $\rho(\mathbf{A}) \leqslant ||\mathbf{A}||$.

En général, $\rho(\mathbf{A}) \neq ||\mathbf{A}||$. Cependant lorsque $\mathbf{A}$ est une matrice hermitienne alors $\rho(\mathbf{A}) = ||\mathbf{A}||_2$ car, d'après (1.20) :

$$||\mathbf{A}||_2 = \left(\rho(\mathbf{A}^*\mathbf{A})\right)^{1/2} \quad \text{et} \quad \lambda_i(\mathbf{A}^*\mathbf{A}) = \left(\lambda_i(\mathbf{A})\right)^2 \tag{1.24}$$

Théorème 1.17 Soit $\mathbf{A} \in \mathbb{K}^{n,n}$. Pour tout $\varepsilon > 0$, il existe une norme matricielle $||\cdot||$ subordonnée à une norme vectorielle telle que $||\mathbf{A}|| \leqslant \rho(\mathbf{A}) + \varepsilon$.

1.2 Suites dans $\mathbb{K}^{n,n}$

▷ **Définition 1.18** Soit $\mathbf{A}^{(k)} = (a_{ij}^{(k)})$, $k \in \mathbb{N}$, une suite de matrices de $\mathbb{K}^{n,n}$. On dit que la suite $(\mathbf{A}^{(k)})$ converge vers une matrice $\mathbf{A}$ lorsque $k \rightarrow +\infty$ si :

$$\forall i = 1, \ldots, n, \quad \forall j = 1, \ldots, n, \quad \lim_{k\rightarrow+\infty} a_{ij}^{(k)} = a_{ij} \tag{1.25}$$

On considère maintenant une suite particulière.

Théorème 1.19 Soit $\mathbf{A} \in \mathbb{K}^{n,n}$ et considérons la suite $(\mathbf{A}^k)$ de $\mathbb{K}^{n,n}$ définie par $\mathbf{A}^0 = \mathbf{A}$ et $\mathbf{A}^{k+1} = \mathbf{A}\,\mathbf{A}^k$ pour $k \in \mathbb{N}$ alors :

$$\lim_{k\to+\infty} \mathbf{A}^k = \mathbf{0} \quad \Leftrightarrow \quad \rho(\mathbf{A}) < 1 \tag{1.26}$$

▷ **Corollaire 1.20** Si $\mathbf{A} \in \mathbb{K}^{n,n}$ vérifie $||\mathbf{A}|| < 1$ pour une norme vectorielle alors :

$$\lim_{k\to+\infty} \mathbf{A}^k = \mathbf{0} \tag{1.27}$$

2

Résolution numérique de systèmes linéaires

▷ **Problème 2.1** Étant donnés une matrice carrée $\mathbf{A} \in \mathbb{K}^{n,n}$ et un vecteur $\mathbf{b} \in \mathbb{K}^n$, trouver $\mathbf{x} \in \mathbb{K}^n$ vérifiant :

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.1}$$

On distingue deux grandes classes de méthodes de résolution numérique :
- les méthodes directes ;
- les méthodes itératives.

2.1 Méthodes directes

Ces méthodes donnent la solution du problème (lorsqu'elle existe) en un nombre *fini* d'opérations. Elles ramènent la résolution d'un système linéaire quelconque à celle d'un système triangulaire. Pour chaque méthode, il est important de considérer deux aspects :
- le nombre d'opérations arithmétiques nécessitées par la méthode : il est montré, par exemple, que la méthode de Cramer est inutilisable numériquement ;
- l'influence des erreurs d'arrondi sur la précision de la méthode.

2.1.1 Système triangulaire

Lorsque $\mathbf{A}$ est une matrice triangulaire supérieure, le système linéaire à résoudre est :

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \ldots\ldots\ldots & \\ a_{nn}x_n &= b_n \end{aligned} \tag{2.2}$$

Ce système a une solution $\mathbf{x} = (x_i)_{i=1,\ldots,n}$ si et seulement si :

$$\forall i = 1, \ldots, n, \quad a_{ii} \neq 0 \tag{2.3}$$

Lorsque cette condition est vérifiée, on calcule les composantes de $\mathbf{x}$ « en remontant », soit :

$$\begin{aligned} x_n &= \frac{b_n}{a_{nn}} \\ x_i &= \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{nn}}, \quad \forall i = n-1, \ldots, 1 \end{aligned} \tag{2.4}$$

Notons que la condition $a_{ii} \neq 0$ pour $i = 1, \ldots, n$ est une condition nécessaire et suffisante pour que la matrice triangulaire $\mathbf{A}$ soit inversible.

Remarque Numériquement, la condition (2.3) se traduit par :

$$\forall i = 1, \ldots, n, \quad |a_{ii}| > \varepsilon \tag{2.5}$$

pour un $\varepsilon > 0$ qui dépend de la précision des calculs.

Nombre d'opérations

Le nombre d'opérations arithmétiques nécessitées par (2.5) se détaille comme suit : $n(n-1)/2$ additions, $n(n-1)/2$ multiplications et n divisions, soit n^2 opérations élémentaires.

2.1.2 Méthodes de Gauss

Algorithme

Réécrivons le système linéaire (2.1) par :

$$\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)} \quad \text{avec} \quad \mathbf{A}^{(1)} = \mathbf{A} \ \text{ et } \ \mathbf{b}^{(1)} = \mathbf{b} \tag{2.6}$$

Sous une hypothèse précisée plus loin, la première étape élimine x_1 dans les $n-1$ dernières équations, la deuxième étape élimine x_2 dans les $n-2$ dernières équations et ainsi de suite, la p^{e} étape (avec $p \in \{1, \ldots, n-1\}$) élimine x_p dans les $n-p$ dernières équations, jusqu'à l'élimination de x_{n-1}. Supposons qu'à l'étape p avec $p \in \{1, \ldots, n-1\}$, le système linéaire s'écrive :

$$\begin{bmatrix} a_{11}^{(p)} & a_{12}^{(p)} & \cdots & a_{1p}^{(p)} & \cdots & a_{1n}^{(p)} \\ 0 & a_{22}^{(p)} & \cdots & a_{2p}^{(p)} & \cdots & a_{2n}^{(p)} \\ \vdots & \ddots & \ddots & \vdots & & \vdots \\ 0 & & 0 & a_{pp}^{(p)} & \cdots & a_{pn}^{(p)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{np}^{(p)} & \cdots & a_{nn}^{(p)} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^{(p)} \\ b_2^{(p)} \\ \vdots \\ b_p^{(p)} \\ \vdots \\ b_n^{(p)} \end{pmatrix} \tag{2.7}$$

On écrira :

$$\mathbf{A}^{(p)}\mathbf{x} = \mathbf{b}^{(p)} \tag{2.8}$$

Pour passer de l'étape p à l'étape $p+1$:

— si $a_{pp}^{(p)} \neq 0$: on élimine x_p dans les $p-1$ dernières équations en multipliant pour $i = p+1, \ldots, n$, la ligne p par :

$$m_{ip} = a_{ip}^{(p)} / a_{pp}^{(p)} \tag{2.9}$$

et en soustrayant la ligne obtenue à la ligne i. Ceci s'écrit de manière équivalente, pour $i = p+1, \ldots, n$:

$$a_{ij}^{(p+1)} = a_{ij}^{(p)} - m_{ip} a_{pj}^{(p)}, \quad j = p, \ldots, n \tag{2.10}$$

$$b_i^{(p+1)} = b_i^{(p)} - m_{ip} b_p^{(p)} \tag{2.11}$$

Les p premières lignes sont inchangées. Le coefficient $a_{pp}^{(p)}$ est appelé le *pivot*.

— si $a_{pp}^{(p)} = 0$ ou si $a_{pp}^{(p)}$ est très petit : lorsque $\mathbf{A}$ est inversible, il existe toujours (au moins) un indice $i_0 \in \{p+1, \ldots, n\}$ tel que $a_{i_0 p}^{(p)} \neq 0$. On permute alors la ligne p du système linéaire avec une la ligne i_0 (ceci revient à permuter les lignes p et i_0 de $\mathbf{A}$ et $\mathbf{b}$) avant d'appliquer les formules (2.9) à (2.11).

Par conséquent, si pour $p \in \{1, \ldots, n-1\}$, il existe un pivot non nul alors $\mathbf{A}^{(n)}$ est une matrice triangulaire supérieure. La résolution numérique de ce système est présentée en sous-section 2.1.1.

Nombre d'opérations

Le nombre d'opérations nécessaires est $2n^3/3 + \mathcal{O}(n^2)$ qui est à comparer au nombre d'opérations nécessaire dans la méthode de Cramer.

Choix du pivot

Par une étude de propagation d'erreur, on montre que l'on a intérêt à choisir le pivot de module le plus grand possible afin d'améliorer la stabilité de la méthode.

Méthode de Gauss avec pivot total Pour passer de l'étape p à l'étape $p+1$, le pivot choisi est un élément de plus grand module dans la sous-matrice :

$$\bar{\mathbf{A}}^{(p)} = (a_{ij}^{(p)})_{i,j=p,\ldots,n} \tag{2.12}$$

Le pivot choisi est donc un coefficient de la matrice $\mathbf{A}$ d'indices (i_0, j_0) tel que :

$$|a_{i_0 j_0}^{(p)}| = \max_{i,j=p,\ldots,n} |a_{ij}^{(p)}| \tag{2.13}$$

C'est la méthode la plus *stable*, mais, en plus des permutations de lignes de $\mathbf{A}$ et $\mathbf{b}$, elle nécessite des permutations de colonnes de $\mathbf{A}$ et donc un changement de numérotation des inconnues.

Méthode de Gauss avec pivot partiel Pour passer de l'étape p à l'étape $p+1$, le pivot choisi est un élément de plus grand module dans la partie de la colonne p d'indice de ligne supérieur ou égal à p, c'est-à-dire dans l'ensemble $\{a_{ip}^{(p)};\ i = p, \ldots, n\}$. Le pivot choisi est donc un coefficient de la matrice $\mathbf{A}$ d'indices (i_0, p) tel que :

$$|a_{i_0 p}^{(p)}| = \max_{i=p,\ldots,n} |a_{ip}^{(p)}| \tag{2.14}$$

C'est la méthode classique. Elle requiert uniquement des permutations de lignes de $\mathbf{A}$ et $\mathbf{b}$.

Analyse matricielle

▷ **Définition 2.2 — Matrice de permutations.** On appelle matrice de permutations une matrice $\mathbf{P} \in \mathbb{K}^{n,n}$ de coefficients $(p_{ij})_{i,j=1,\ldots,n}$ telle qu'il existe une permutation σ de l'ensemble $\{1, \ldots, n\}$ telle que :

$$p_{ij} = \delta_{\sigma(i)j}, \quad i, j = 1, \ldots, n \tag{2.15}$$

■ **Exemple** Si k et ℓ sont deux entiers fixés dans $\{1, \ldots, n\}$ et si σ est la transposition de $\{1, \ldots, n\}$ qui échange k et ℓ, alors la matrice de permutations correspondante, notée $\mathbf{P}^{(k,\ell)}$:

$$\mathbf{P}^{(k,\ell)} = \begin{array}{c} \\ k \\ \\ \ell \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & k & & \ell & & \end{array} \\ \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 0 & \cdots & 1 & \\ & & & 1 & \ddots & \vdots & \\ & & & \vdots & & 1 & \\ & & & 1 & \cdots & 0 & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{bmatrix} \end{array} \tag{2.16}$$

a les propriétés suivantes :
— le produit $\mathbf{P}^{(k,\ell)}\mathbf{A}$ a pour effet de permuter les lignes k et ℓ de $\mathbf{A}$;
— le produit $\mathbf{A}\mathbf{P}^{(k,\ell)}$ a pour effet de permuter les colonnes k et ℓ de $\mathbf{A}$.
pour toute matrice $\mathbf{A}$ de $\mathbb{K}^{n,n}$. La matrice $\mathbf{P} \equiv \mathbf{P}^{(k,\ell)}$ est inversible et $\mathbf{P}^{-1} = \mathbf{P}^{\top} = \mathbf{P}$:

Lorsque la matrice d'un système linéaire est inversible, on peut toujours utiliser la méthode de Gauss avec pivot partiel. Ce résultat s'écrit matriciellement de la façon suivante :

Théorème 2.3 Soit $\mathbf{A} \in \mathbb{K}^{n,n}$ une matrice inversible, alors il existe une matrice de permutations $\mathbf{P} \in \mathbb{K}^{n,n}$ telle que le produit $\mathbf{PA}$ se décompose en :

$$\mathbf{PA} = \mathbf{LU} \tag{2.17}$$

où $\mathbf{L}$ est une matrice triangulaire inférieure *ayant des 1 sur la diagonale* et où $\mathbf{U}$ est une matrice triangulaire supérieure.

La factorisation $\mathbf{LU}$ de $\mathbf{PA}$ est unique si on impose des 1 sur la diagonale de $\mathbf{L}$. Elle permet de résoudre un système linéaire en résolvant deux systèmes linéaires triangulaires :

$$\mathbf{Ax} = \mathbf{b} \quad \Leftrightarrow \quad \mathbf{PAx} = \mathbf{Pb} \quad \Leftrightarrow \quad \mathbf{LUx} = \mathbf{Pb} \quad \Leftrightarrow \quad \mathbf{Ly} = \mathbf{Pb} \text{ et } \mathbf{Ux} = \mathbf{y} \tag{2.18}$$

Lorsque $\mathbf{P} = \mathbf{I}$ (algorithme sans permutation), l'algorithme de Gauss montre que la matrice $\mathbf{U}$ est alors la matrice $\mathbf{A}^{(n)}$. Elle est aussi une méthode de calcul de déterminants. En effet, il résulte du théorème 2.3 que :

▷ **Proposition 2.4** Si la matrice $\mathbf{A}$ est inversible et sous les notations précédentes :

$$\det(\mathbf{A}) = (-1)^{n_{\text{per}}} \prod_{p=1}^{n} a_{pp}^{(p)} \tag{2.19}$$

où n_{per} est le nombre de permutations de lignes ou de colonnes effectuées dans un algorithme de Gauss avec pivot partiel.

En effet, $\det(\mathbf{A}) = \det(\mathbf{P}^{-1})\det(\mathbf{L})\det(\mathbf{U})$ avec $\det(\mathbf{L}) = 1$ et $\det(\mathbf{U}) = \prod_{i=1}^{n} a_{pp}^{(p)}$ puisque $\mathbf{L}$ et $\mathbf{U}$ sont des matrices triangulaires.

La méthode de Gauss est une méthode directe de résolution de systèmes linéaires utilisable dès que la matrice $\mathbf{A}$ du système est inversible. Elle présente le défaut de propager facilement les erreurs d'arrondis. On lui préférera donc d'autres méthodes lorsque n est grand ou que la matrice est *mal conditionnée*.

2.1.3 Méthode LU ou algorithme de Crout

Lorsque la décomposition $\mathbf{A} = \mathbf{LU}$ du théorème 2.3 est possible sans permutation, il est parfois préférable de calculer les coefficients de $\mathbf{L}$ et ceux de $\mathbf{U}$ puis de résoudre les deux systèmes triangulaires :

$$\begin{cases} \mathbf{Ly} = \mathbf{b} \\ \mathbf{Ux} = \mathbf{y} \end{cases} \tag{2.20}$$

où $\mathbf{x}$ et $\mathbf{y} \in \mathbb{K}^n$. Les coefficients de $\mathbf{L}$ et de $\mathbf{U}$ sont donnés par :
— pour $i = 2, \ldots, n$ et $j = 1, \ldots, i-1$: $\ell_{ji} = u_{ij} = 0$;
— pour $i = 1, \ldots, n$ et $j = i, \ldots, n$:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} \ell_{ik} u_{kj} \quad ; \quad \ell_{ji} = \frac{a_{ij} - \sum_{k=1}^{i-1} \ell_{jk} u_{ki}}{u_{ii}} \tag{2.21}$$

2.1.4 Méthode de Cholesky

Cette méthode est utilisée pour les matrices réelles, symétriques définies positives. On suppose dans tout ce paragraphe que $\mathbf{A}$ est une matrice réelle, $\mathbf{A} \in \mathbb{R}^{n,n}$.

▷ **Proposition 2.5** Une condition nécessaire et suffisante pour qu'une matrice $\mathbf{A} \in \mathbb{R}^{n,n}$ soit symétrique définie positive est qu'il existe une matrice $\mathbf{S} \in \mathbb{R}^{n,n}$ triangulaire inférieure et inversible telle que :

$$\mathbf{A} = \mathbf{S}\mathbf{S}^\top \tag{2.22}$$

Démonstration. La démonstration donne par récurrence les coefficients de la matrice $\mathbf{S}$.

Condition suffisante Supposons l'existence de $\mathbf{S}$ alors, pour tout $\mathbf{x} \neq 0$, $\mathbf{x} \in \mathbb{R}^n$:

$$(\mathbf{A}\mathbf{x}, \mathbf{x}) = (\mathbf{S}\mathbf{S}^\top \mathbf{x}, \mathbf{x}) = (\mathbf{S}^\top \mathbf{x}, \mathbf{S}^\top \mathbf{x}) = ||\mathbf{S}^\top \mathbf{x}||^2 > 0 \tag{2.23}$$

dès que $\mathbf{S}$ est inversible. Pour tout $\mathbf{x} \in \mathbb{R}^n$ et tout $\mathbf{y} \in \mathbb{R}^n$:

$$(\mathbf{A}\mathbf{x}, \mathbf{y}) = (\mathbf{S}^\top \mathbf{x}, \mathbf{S}^\top \mathbf{y}) = (\mathbf{x}, \mathbf{A}\mathbf{y}) \tag{2.24}$$

donc $\mathbf{A}$ est symétrique.

Condition nécessaire On construit une matrice $\mathbf{S}$ qui répond à la question. Si $\mathbf{S} = (s_{ij})_{i,j=1,\dots,n}$ existe alors les coefficients s_{ij} doivent vérifier $\forall i, j = 1, \dots, n$:

$$a_{ij} = \sum_{k=1}^{n} s_{ik} s_{kj}^\top = \sum_{k=1}^{n} s_{ik} s_{jk} \tag{2.25}$$

$$s_{ij} = 0 \quad \text{si } j > i \tag{2.26}$$

donc pour $j \geqslant i$:

$$a_{ij} = \sum_{k=1}^{n} s_{ik} s_{jk} \tag{2.27}$$

On utilise le lemme suivant :

▷ **Lemme 2.6** Soit $\mathbf{A}$ une matrice réelle symétrique définie positive, alors :

$$\forall k = 1, \dots, n, \quad \det \mathbf{A}_k > 0 \tag{2.28}$$

où $\mathbf{A}_k = (a_{ij})_{i,j=1,\dots,k} \in \mathbb{R}^{k,k}$ et les matrices $\mathbf{A}_k$ sont symétriques définies positives.

Démonstration (suite). On détermine par récurrence les coefficients de la colonne i de $\mathbf{S}$.

— Lorsque $i = 1$:

— $j = 1$: la condition (2.27) implique $a_{11} = s_{11}^2$. Le lemme 2.6 avec $k = 1$ implique $\det \mathbf{A}_1 = a_{11} > 0$. On peut donc choisir :

$$s_{11} = \sqrt{a_{11}} > 0 \tag{2.29}$$

— $j > 1$: la condition (2.27) implique $a_{1j} = s_{11} s_{1j}$ donc :

$$s_{j1} = \frac{a_{1j}}{a_{11}} \quad j = 2, \dots, n \tag{2.30}$$

— Pour $i = 2, \dots, n$: supposons les $(i-1)$ premières colonnes de $\mathbf{S}$ connues avec $s_{mm} > 0$ pour $m = 1, \dots, i-1$ et $s_{jm} = 0$ pour $j < m$, alors d'après (2.25) à (2.27) :

— si $j = i$ alors :

$$a_{ii} = s_{ii}^2 + \sum_{k=1}^{i-1} s_{ik}^2 \tag{2.31}$$

Il résulte du lemme 2.6 que :

$$a_{ii} = \sum_{k=1}^{i-1} s_{ik}^2 > 0 \tag{2.32}$$

On peut donc choisir :

$$s_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} s_{ik}^2} \tag{2.33}$$

— si $j > i$ alors :

$$s_{ji} = \frac{a_{ij} - \sum_{k=1}^{i-1} s_{ik} s_{jk}}{s_{ii}} \tag{2.34}$$

II en résulte la méthode suivante de résolution d'un système linéaire :

▷ **Corollaire 2.7 — Méthode de Cholesky.** Lorsque $\mathbf{A}$ est une matrice réelle, symétrique, définie positive, résoudre le système linéaire (2.1) revient à résoudre successivement deux systèmes linéaires triangulaires :

$$\begin{cases} \mathbf{S}\mathbf{y} = \mathbf{b} \\ \mathbf{S}^\top \mathbf{x} = \mathbf{y} \end{cases} \tag{2.35}$$

où $\mathbf{S} = (s_{ij})$ est la matrice triangulaire inférieure donnée par l'algorithme : pour $i = \ell, \ldots, n$

$$s_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} s_{ik}^2} \tag{2.36}$$

$$s_{ji} = \frac{a_{ij} - \sum_{k=1}^{i-1} s_{ik} s_{jk}}{s_{ii}} \quad \text{si } j = i+1, \ldots, n \tag{2.37}$$

$$s_{ji} = 0 \quad \text{si } j < i \tag{2.38}$$

Remarques

1. Si la matrice symétrique $\mathbf{A}$ n'est pas définie positive, il existe un indice i tel que :
$$a_{ii} - \sum_{k=1}^{i-1} s_{ik}^2 \leqslant 0 \tag{2.39}$$
La condition inverse :
$$a_{ii} - \sum_{k=1}^{i-1} s_{ik}^2 > 0 \tag{2.40}$$
donne donc un test permettant de vérifier si une matrice $\mathbf{A}$ est ou n'est pas définie positive.
2. La décomposition de $\mathbf{A}$ en $\mathbf{S}\mathbf{S}^\top$ est unique si on choisit les éléments diagonaux de $\mathbf{S}$ (strictement) positifs.
3. Si la matrice $\mathbf{A}$ est une matrice $(2\ell+1)$ diagonale, alors la matrice $\mathbf{S}$ est une matrice bande $(\ell+1, 1)$.
4. La méthode de Cholesky est une méthode de calcul de déterminant :
$$\det \mathbf{A} = \prod_{i=1}^{n} s_{ii}^2 \tag{2.41}$$
5. La méthode de Cholesky est moins sensible aux erreurs d'arrondi que la méthode de Gauss (si la racine carrée est évaluée correctement).

Nombre d'opérations

Il y a $n^3/3 + \mathcal{O}(n^2)$ opérations élémentaires qu'il est intéressant de comparer à la méthode de Gauss quand la matrice du système est symétrique. La méthode de Cholesky est donc une excellente méthode à privilégier devant la méthode de Gauss lorsque la matrice du système est réelle symétrique définie positive.

2.2 Méthodes itératives

Une classe importante de méthodes de résolution de systèmes linéaires est celle des méthodes itératives qui donnent la solution du système comme limite d'une suite de vecteurs.

2.2.1 Résultats généraux

On supposera dans cette section que $\mathbf{A} \in \mathbb{K}^{n,n}$ est une matrice inversible. On se propose de construire une suite $(\mathbf{x}^{(m)})_{m\in\mathbb{N}}$ d'éléments de $\mathbb{K}^n$ telle qu'il existe :

$$\lim_{m\to+\infty} \mathbf{x}^{(m)} = \mathbf{x} \tag{2.42}$$

avec $\mathbf{x}$ vérifiant l'équation (2.1). Les schémas étudiés seront de la forme :

$$\begin{cases} \mathbf{x}^{(0)} \text{ donné dans } \mathbb{K}^n \\ \mathbf{x}^{(m+1)} = \mathbf{B}\mathbf{x}^{(m)} + \mathbf{c} \end{cases} \tag{2.43}$$

où $\mathbf{B} \in \mathbb{K}^{n,n}$ et $\mathbf{c} \in \mathbb{K}^n$ sont donnés.

▷ **Définition 2.8 — Convergence.** Une méthode itérative de la forme (2.43) est dite convergente si pour tout vecteur initial $\mathbf{x}^{(0)}$, $\mathbf{x}^{(m)}$ admet une limite quand $m \to \infty$ et $\lim \mathbf{x}^{(m)}$ (quand $m \to \infty$) est solution de (2.1).

Construction de méthodes itératives La matrice $\mathbf{A}$ est décomposée en $\mathbf{A} = \mathbf{M} - \mathbf{N}$ où $\mathbf{M}$ est une matrice « facile à inverser ». Le système à résoudre (2.1) s'écrit alors :

$$\mathbf{M}\mathbf{x} = \mathbf{N}\mathbf{x} + \mathbf{b} \tag{2.44}$$

On définit un schéma itératif par :

$$\begin{cases} \mathbf{x}^{(0)} \text{ donné dans } \mathbb{K}^n \\ \mathbf{M}\mathbf{x}^{(m+1)} = \mathbf{N}\mathbf{x}^{(m)} + \mathbf{b} \end{cases} \tag{2.45}$$

Théorème 2.9 La méthode itérative (2.45) est convergente si et seulement si :

$$\rho(\mathbf{M}^{-1}\mathbf{N}) < 1 \tag{2.46}$$

Démonstration. On définit $\mathbf{e}^{(m)} = \mathbf{x}^{(m)} - \mathbf{x}$ l'erreur d'approximation à la m^{e} étape et on a :

$$\mathbf{e}^{(m)} = \mathbf{M}^{-1}\mathbf{N}\mathbf{e}^{(m-1)} = (\mathbf{M}^{-1}\mathbf{N})^m \mathbf{e}^{(0)} \tag{2.47}$$

puis on utilise le théorème 1.19.

▷ **Proposition 2.10** Le nombre d'itérations m_0 réduisant l'erreur d'approximation d'un facteur η dans une méthode itérative de la forme (2.43) vérifie :

$$m_0 \geqslant \left| \frac{\ln(\eta)}{\ln(\rho(\mathbf{B}))} \right| \tag{2.48}$$

Ce résultat nous donne un test de comparaison de deux méthodes itératives.

▷ **Définition 2.11** On dit qu'une méthode itérative définie par une matrice $\mathbf{B}_1$ est asymptotiquement plus rapide que celle définie par une matrice $\mathbf{B}_2$ si :

$$\rho(\mathbf{B}_1) \leqslant \rho(\mathbf{B}_2) \tag{2.49}$$

2.2.2 Principales méthodes itératives

Soit $\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F}$ la décomposition de $\mathbf{A}$ définie par :

$$\mathbf{D} = (d_{ij})_{i,j=1,\ldots,n} \quad \text{avec} \quad \begin{cases} d_{ii} = a_{ii} & i = 1,\ldots,n \\ d_{ij} = 0 & i \neq j \end{cases} \tag{2.50}$$

$$\mathbf{E} = (e_{ij})_{i,j=1,\ldots,n} \quad \text{avec} \quad \begin{cases} e_{ij} = -a_{ij} & i > j \\ e_{ij} = 0 & i \leqslant j \end{cases} \tag{2.51}$$

$$\mathbf{F} = (f_{ij})_{i,j=1,\ldots,n} \quad \text{avec} \quad \begin{cases} f_{ij} = -a_{ij} & i < j \\ f_{ij} = 0 & i \geqslant j \end{cases} \tag{2.52}$$

Méthode de Jacobi

On pose :

$$\begin{cases} \mathbf{M} = \mathbf{D} \\ \mathbf{N} = \mathbf{E} + \mathbf{F} \end{cases} \tag{2.53}$$

La matrice d'itérations s'écrit :

$$\mathcal{J} = \mathbf{B} = \mathbf{D}^{-1}(\mathbf{E} + \mathbf{F}) \tag{2.54}$$

et la méthode nécessite $a_{ii} \neq 0$ pour $i = 1,\ldots,n$. Les composantes du vecteur $\mathbf{x}^{(m+1)}$ sont données par :

$$a_{ii} x_i^{(m+1)} = -\sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(m)} + b_i, \quad i = 1,\ldots,n \tag{2.55}$$

Remarque Il est nécessaire de conserver le vecteur $\mathbf{x}^{(m)}$ tout entier en mémoire pour calculer $\mathbf{x}^{(m+1)}$.

Méthode de Gauss-Seidel

On pose

$$\begin{cases} \mathbf{M} = \mathbf{D} - \mathbf{E} \\ \mathbf{N} = \mathbf{F} \end{cases} \tag{2.56}$$

La matrice d'itérations est :

$$\mathcal{L} = \mathbf{B} = (\mathbf{D} - \mathbf{E})^{-1}\mathbf{F} \tag{2.57}$$

et la méthode nécessite $a_{ii} \neq 0$ pour $i = 1,\ldots,n$. Les composantes du vecteur $\mathbf{x}^{(m+1)}$ sont données par :

$$\sum_{j=1}^{i} a_{ij} x_j^{(m+1)} = -\sum_{j=i+1}^{n} a_{ij} x_j^{(m)} + b_i, \quad i = 1,\ldots,n \tag{2.58}$$

Méthode de relaxation de paramètre ω
Soit ω un paramètre réel non nul. On pose :

$$\begin{cases} \mathbf{M} = \mathbf{D}/\omega - \mathbf{E} \\ \mathbf{N} = (1/\omega - 1)\mathbf{D} + \mathbf{F} \end{cases} \tag{2.59}$$

La matrice d'itérations s'écrit :

$$\mathcal{L}_\omega = \mathbf{B} = \left(\frac{1}{\omega}\mathbf{D} - \mathbf{E}\right)^{-1}\left(\left(\frac{1}{\omega} - 1\right)\mathbf{D} + \mathbf{F}\right) = (\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{E})^{-1}\left((1-\omega)\mathbf{I} + \omega\mathbf{D}^{-1}\mathbf{F}\right) \tag{2.60}$$

La méthode nécessite $a_{ii} \neq 0$ pour $i = 1, \ldots, n$ et les composantes du vecteur $\mathbf{x}^{(m+1)}$ sont données pour $i = 1, \ldots, n$ par :

$$a_{ii}x_i^{(m+1)} = -\omega \sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} + (1-\omega)a_{ii}x_i^{(m)} - \omega \sum_{j=i+1}^{n} a_{ij}x_j^{(m)} + \omega b_i \tag{2.61}$$

— si $\omega < 1$, la méthode est dite de *sous-relaxation* ;
— si $\omega = 1$, on retrouve la méthode de Gauss-Seidel ;
— si $\omega > 1$, la méthode est dite de *sur-relaxation*.

Un problème important est celui de la détermination « optimale » du paramètre ω.

2.3 Convergence et comparaison des méthodes itératives

Une des principales difficultés des méthodes itératives est le problème de la convergence des suites $(\mathbf{x}^{(k)})$. Donnons quelques cas classiques de convergence des méthodes décrites plus haut.

2.3.1 Matrices à diagonale strictement dominante

Reprenons la définition 1.10 : on rappelle qu'une matrice à diagonale strictement dominante est inversible.

Théorème 2.12 Si $\mathbf{A}$ est une matrice à diagonale strictement dominante, les méthodes itératives de Jacobi et de Gauss-Seidel sont convergentes.

Démonstration. Voir l'exercice 13.

2.3.2 Matrices hermitiennes définies positives

▷ **Proposition 2.13** Une condition nécessaire pour qu'une méthode de relaxation converge est que :

$$0 < \omega < 2 \tag{2.62}$$

La démonstration utilise le lemme suivant :

▷ **Lemme 2.14** Pour tout $\omega \in \mathbb{R}$, nous avons :

$$\rho(\mathcal{L}_\omega) \geqslant |\omega - 1| \tag{2.63}$$

avec égalité si et seulement si toutes les valeurs propres de $\mathcal{L}_\omega$ ont pour module $|\omega - 1|$.

Théorème 2.15 Supposons $\mathbf{A}$ hermitienne, définie positive. Si $0 < \omega < 2$, les méthodes de relaxation de paramètre ω convergent.

Ce résultat est vrai en particulier pour la méthode de Gauss-Seidel.

▷ **Proposition 2.16** Soit $\mathbf{A}$ une matrice hermitienne, définie positive et tridiagonale, alors les méthodes de Jacobi, de Gauss-Seidel et de relaxation avec $0 < \omega < 2$ sont convergentes. De plus, il existe un paramètre ω_0, optimal tel que :

$$\begin{cases} \rho(\mathcal{L}_{\omega_0}) \leqslant \rho(\mathcal{L}_\omega), \quad \forall \omega \in]0\,;2[\\ \rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_1) < \rho(\mathcal{J}) < 1 \\ 1 < \omega_0 < 2 \end{cases} \tag{2.64}$$

Remarques Il faut attacher de l'importance aux aspects suivants :

1. critères d'arrêt utilisés ;
2. estimation d'un paramètre « optimal » ;
3. ces méthodes sont peu sensibles aux erreurs d'arrondi.

Ces méthodes sont particulièrement intéressantes pour les grands systèmes ou lorsque la matrice $\mathbf{A}$ a une structure particulière (à diagonale strictement dominante ou matrice creuse) à condition que la convergence soit assurée.

Calcul de valeurs et vecteurs propres

▷ **Problème 3.1** Étant donnée une matrice $\mathbf{A} \in \mathbb{R}^{n,n}$, on cherche les (ou des) valeurs propres λ de $\mathbf{A}$ et des vecteurs propres correspondants.

On distinguera deux types de méthodes :

— les méthodes de la puissance itérée ;

— les méthodes basées sur des transformations matricielles.

Rappels

$\lambda \in \mathbb{C}$ est une valeur propre de $\mathbf{A}$ s'il existe $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x} \neq \mathbf{0}$, vérifiant :

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x} \tag{3.1}$$

— Un vecteur $\mathbf{x} \neq \mathbf{0}$ vérifiant (3.1) est un vecteur propre (à droite) de $\mathbf{A}$ associé à la valeur propre λ.

— Un vecteur $\mathbf{y} \in \mathbb{C}^n$, $\mathbf{y} \neq \mathbf{0}$ vérifiant $\mathbf{y}^\top \mathbf{A} = \lambda \mathbf{y}^\top$ est appelé vecteur propre à gauche de $\mathbf{A}$ associé à λ.

▷ **Proposition 3.2**

1. Si $\mathbf{u}_1, \ldots, \mathbf{u}_p$ $(p \leqslant n)$ sont p vecteurs propres de $\mathbf{A}$ associés à p valeurs propres distinctes, ils sont indépendants.
2. Si λ est une valeur propre de $\mathbf{A}$, il existe un vecteur propre à gauche de $\mathbf{A}$. Ce vecteur propre à gauche est alors un vecteur propre à droite de $\mathbf{A}^\top$ associé à λ. Si $\mathbf{A}$ est symétrique, un vecteur propre à gauche est aussi un vecteur propre à droite.
3. Si $\mathbf{y}$ est un vecteur propre à gauche de $\mathbf{A}$ et $\mathbf{x}$, un vecteur propre à droite de $\mathbf{A}$ associés à des valeurs propres distinctes, alors $\mathbf{y}^\top \cdot \mathbf{x} = 0$.

3.1 Vecteurs propres d'une matrice triangulaire

▷ **Proposition 3.3** Soit $\mathbf{A}$ une matrice triangulaire supérieure. Si les valeurs propres $\lambda_i = a_{ii}$, $i = 1, \ldots, n$ sont toutes distinctes alors les vecteurs propres $\mathbf{u}_i$ associés s'écrivent :

— pour $i = 1$: $\mathbf{u}_1 = (1, 0, \ldots, 0)^\top$

— pour $i = 2, \ldots, n$, si on note $\mathbf{u}_i = (u_{ji})_{j=1,\ldots,n}$:

$$\begin{cases} u_{ji} = 0, & j = i+1, \ldots, n \\ u_{ii} = 1 \\ u_{ji} = -\dfrac{a_{ji} + \sum_{k=j+1}^{i-1} a_{jk} u_{ki}}{\lambda_j - \lambda_i}, & j = i-1, \ldots, 1 \quad \text{(pas } -1\text{)} \end{cases} \tag{3.2}$$

à un coefficient multiplicatif près

Démonstration. Voir l'exercice 16 dans le cas d'une matrice triangulaire inférieure.

3.2 Méthodes de la puissance itérée

Ce paragraphe décrit un premier type de méthodes utilisées généralement pour le calcul d'une ou de quelques valeurs propres d'une matrice, ainsi que des vecteurs propres associés.

3.2.1 Itérations simples

Théorème 3.4 Soit $\mathbf{A} \in \mathbb{R}^{n,n}$. On suppose que la valeur propre de plus grand module est simple et vérifie $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$ et que $\mathbf{A}$ possède n vecteurs propres $\mathbf{u}_1, \ldots, \mathbf{u}_n$ linéairement indépendants, autrement dit $\mathbf{A}$ est diagonalisable. Soit $\|\cdot\|$ une norme vectorielle. On construit une suite $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^\top$ d'éléments de $\mathbb{R}^n$ par récurrence :

$$\begin{cases} \mathbf{x}^{(0)} \text{ donné dans } \mathbb{R}^n \\ \mathbf{x}^{(k+1)} = \dfrac{\mathbf{A}\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|}, \quad k \in \mathbb{N} \end{cases} \tag{3.3}$$

Si $\mathbf{x}^{(0)}$ n'est pas orthogonal à l'espace propre à gauche associé à λ_1, alors :

1. la suite $((\lambda_1/|\lambda_1|)^k \mathbf{x}^{(k)})$ converge vers un vecteur propre associé à λ_1 quand $k \to +\infty$;
2. $\lim_{k\to+\infty} (\mathbf{A}\mathbf{x}^{(k)})_j / x_j^{(k)} = \lambda_1$ pour au moins un $j \in \{1, \ldots, n\}$ où on a noté $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^\top$.

On choisit en général la norme $\|\cdot\|_\infty$.

Démonstration. On procède en deux étapes :

1. Soit $\mathbf{A}$ diagonalisable dans $\mathbb{C}$. On note $\lambda_1, \ldots, \lambda_n$ les valeurs propres de $\mathbf{A}$ dans $\mathbb{C}$, et $\mathbf{u}_1, \ldots, \mathbf{u}_n$, ses vecteurs propres dans $\mathbb{C}^n$. On suppose $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$. Ceci implique que $\lambda_1 \in \mathbb{R}$: en effet, si λ_1 est valeur propre, $\bar{\lambda}_1$ l'est aussi et $|\lambda_1| = |\bar{\lambda}_1|$ ce qui contredit $|\lambda_1| > |\lambda_2|$.
   Soit $\mathbf{x}^{(0)}$ non orthogonal à $\mathbf{v}_1$ défini comme suit :
   $$\mathbf{x}^{(0)} = \sum_{i=1}^{n} \alpha_i \mathbf{u}_i \tag{3.4}$$
   où $\mathbf{v}_1$ est valeur propre à gauche de $\mathbf{A}$, $\mathbf{A}^\top \mathbf{v}_1 = \lambda_1 \mathbf{v}_1$. Alors $\alpha_1 \neq 0$ puisque :
   $$(\mathbf{x}^{(0)}, \mathbf{v}_1) = \sum_{i=1}^{n} \alpha_i (\mathbf{u}_i, \mathbf{v}_1) = \alpha_1 (\mathbf{u}_1, \mathbf{v}_1) \tag{3.5}$$
   où $(\mathbf{u}_1, \mathbf{v}_1) \neq 0$. Pour le reste, $(\mathbf{u}_i, \mathbf{v}_1) = 0$ pour $i = 2, \ldots, n$ puisqu'en effet $\lambda_i (\mathbf{u}_i, \mathbf{v}_1) = (\mathbf{A}\mathbf{u}_i, \mathbf{v}_1) = (\mathbf{u}_i, \mathbf{A}^\top \mathbf{v}_1) = \lambda_1 (\mathbf{u}_i, \mathbf{v}_1)$ d'où $(\mathbf{u}_i, \mathbf{v}_1) = 0$ si $\lambda_i \neq \lambda_1$ (sinon $\mathbf{v}_1 = \mathbf{0}$).
2. On calcule la suite $\mathbf{x}^{(k)}$ définie par $\mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{x}^{(k)}$. Autrement dit :
   $$\mathbf{x}^{(k)} = \mathbf{A}^k \mathbf{x}^{(0)} = \mathbf{A}^k \sum_{i=1}^{n} \alpha_i \mathbf{u}_i = \sum_{i=1}^{n} \alpha_i \mathbf{A}^k \mathbf{u}_i = \sum_{i=1}^{n} \alpha_i \lambda_i^k \mathbf{u}_i \tag{3.6}$$
   donc :
   $$\frac{\mathbf{x}^{(k)}}{\lambda_1^k} = \alpha_1 \mathbf{u}_1 + \sum_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{u}_i \quad \text{où} \quad \lim_{k\to\infty} \left(\frac{\lambda_i}{\lambda_1}\right)^k = 0 \quad \text{puisque } |\lambda_i| < |\lambda_1| \tag{3.7}$$
   Finalement :
   $$\lim_{k\to\infty} \frac{\mathbf{x}^{(k)}}{\lambda_1^k} = \alpha_1 \mathbf{u}_1 \tag{3.8}$$
   qui est un vecteur propre associé à λ_1.

■ **Exemple** Soit :

$$\mathbf{A} = \begin{bmatrix} -4 & -6 \\ -6 & -4 \end{bmatrix} \tag{3.9}$$

qui a pour paires propres $[\lambda_1 = -10, \mathbf{u}_1 = (1,1)^\top]$ et $[\lambda_2 = 2, \mathbf{u}_2 = (1,-1)^\top]$. Partant de $\mathbf{x}^{(0)} = (2,1)^\top$, on calcule $\mathbf{x}^{(1)} = (-14,-16)^\top, \ldots, \mathbf{x}^{(4)} = (-15008,-14992)^\top$ et l'on observe que $\lim_{k\to\infty} \mathbf{x}^{(k)}/(-10)^k = (-1{,}5,-1{,}5)^\top$.

Remarques

1. La suite $\mathbf{x}^{(k)}$ est normée à chaque étape pour éviter que les composantes du vecteur ne deviennent trop grandes ou trop petites.
2. Le facteur de convergence vers la première valeur propre est en $\mathcal{O}(|\lambda_2/\lambda_1|)$. La convergence est donc d'autant plus rapide que $|\lambda_1|$ et $|\lambda_2|$ sont distants l'un de l'autre.
3. Il y a encore convergence vers la première valeur propre et un vecteur propre correspondant lorsque la valeur propre $|\lambda_1|$ est multiple et vérifie $\lambda_1 = \lambda_2 = \ldots = \lambda_p$ et $|\lambda_p| > |\lambda_{p+1}|$.

Pour calculer les quelques valeurs propres suivantes, on utilise une méthode de déflation.

Théorème 3.5 — Méthode de déflation. Sous les hypothèses du théorème 3.4 et si $\mathbf{y}_1$ est un vecteur propre à gauche de $\mathbf{A}$ associé à λ_1 alors la matrice :

$$\mathbf{B} = \mathbf{A} - \frac{\lambda_1}{\mathbf{y}_1^\top \mathbf{u}_1} \mathbf{u}_1 \mathbf{y}_1^\top \tag{3.10}$$

a pour valeurs propres :

$$0, \lambda_2, \ldots, \lambda_n \tag{3.11}$$

associées respectivement aux vecteurs propres :

$$\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n \tag{3.12}$$

Les théorèmes 3.4 et 3.5 permettent théoriquement de calculer toutes les valeurs propre d'une matrice $\mathbf{A}$ pour laquelle $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$. Pratiquement, ce procédé est limité au calcul de quelques valeurs propres en raison de l'accumulation des erreurs numériques.

3.2.2 Méthode d'accélération de convergence

Théorème 3.6 — Méthode du quotient de Rayleigh. Sous les hypothèses du théorème 3.4 et si de plus $\mathbf{A}$ est symétrique, on définit une suite de réels R_k par :

$$\begin{cases} \mathbf{x}^{(0)} \text{ donné dans } \mathbb{R}^n \\ \mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{x}^{(k)} \\ \mathrm{R}_k = \dfrac{\mathbf{x}^{(k)\top}\mathbf{A}\mathbf{x}^{(k)}}{\mathbf{x}^{(k)\top}\mathbf{x}^{(k)}} = \dfrac{\mathbf{x}^{(k)\top}\mathbf{x}^{(k+1)}}{\|\mathbf{x}^{(k)}\|_2^2} \end{cases} \tag{3.13}$$

Si $\mathbf{x}^{(0)}$ n'est pas orthogonal au sous-espace propre à gauche associé à $\mathbf{u}_1$ alors la suite (R_k) converge et $\lim_{k\to\infty} \mathrm{R}_k = \lambda_1$. La vitesse de convergence est en $\mathcal{O}(|\lambda_2/\lambda_1|^2)$.

Démonstration. Voir l'exercice 17.

3.2.3 Méthode de la puissance itérée inverse

Elle est utilisée pour déterminer la plus petite valeur propre en module d'une matrice diagonalisable lorsque cette valeur propre est distincte en module des autres valeurs propres. Au lieu d'appliquer la méthode de la puissance itérée à la matrice inverse $\mathbf{A}^{-1}$, on décompose $\mathbf{A}$ en un produit $\mathbf{LU}$, puis on remplace la suite $\mathbf{x}^{(k)}$ donnée en (3.3) par la suite :

$$\begin{cases} \mathbf{x}^{(0)} \text{ donné dans } \mathbb{R}^n \\ \mathbf{L}\mathbf{y} = \mathbf{x}^{(k)} \\ \mathbf{U}\mathbf{x}^{(k+1)} = \mathbf{y} \end{cases} \tag{3.14}$$

Ce procédé est plus stable.

Cette méthode a plusieurs applications :

— recherche d'un vecteur propre associé à une valeur propre connue ;

— recherche de la valeur propre λ « la plus proche » d'un nombre μ donné et calcul d'un vecteur propre associé à λ.

3.3 Méthodes issues de transformations matricielles

Elles donnent, en général, toutes les valeurs propres d'une matrice $\mathbf{A}$.

Rappels Deux matrices $\mathbf{A}$ et $\mathbf{B} \in \mathbb{R}^{n,n}$ sont dites *semblables* s'il existe une matrice $\mathbf{P}$ inversible, telle que :

$$\mathbf{A} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P}, \quad \mathbf{P} \in \mathbb{R}^{n,n} \tag{3.15}$$

Deux matrices semblables ont les mêmes valeurs propres avec le même ordre de multiplicité. Si $\mathbf{A}$ et $\mathbf{B}$ sont semblables avec $\mathbf{A} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P}$ et si $\mathbf{v}$ est un vecteur propre de $\mathbf{B}$ alors $\mathbf{u} = \mathbf{P}^{-1}\mathbf{v}$ est un vecteur propre de $\mathbf{A}$ associé à la même valeur propre.

3.3.1 Méthode de Rutishauser ou du LU − LR

Cette méthode est basée sur la décomposition $\mathbf{LU}$ d'une matrice $\mathbf{A}$. Nous supposons que $\mathbf{A}$ est inversible et se décompose en $\mathbf{A} = \mathbf{LU}$ sans permutations de lignes ou de colonnes.

Algorithme On construit par récurrence, une suite de matrices $(\mathbf{A}_p)$ semblables à $\mathbf{A}$, telles que :

$$\begin{cases} \mathbf{A}_0 \equiv \mathbf{A} \\ \quad = \mathbf{L}_0\mathbf{U}_0 & \text{par décomposition} \\ \mathbf{A}_1 \equiv \mathbf{U}_0\mathbf{L}_0 & \text{par définition} \\ \quad \cdots\cdots\cdots \\ \mathbf{A}_p = \mathbf{L}_p\mathbf{U}_p \\ \mathbf{A}_{p+1} \equiv \mathbf{U}_p\mathbf{L}_p & p \in \mathbb{N} \end{cases} \tag{3.16}$$

où $\forall p \in \mathbb{N}$, $\mathbf{L}_p$ est une matrice triangulaire inférieure ayant des 1 sur la diagonale ; $\mathbf{U}_p$ est une matrice triangulaire supérieure.

Théorème 3.7 Par hypothèse $\mathbf{A}$ est inversible, les valeurs propres de $\mathbf{A}$ sont réelles et distinctes en valeur absolue et les factorisations $\mathbf{L}_p\mathbf{U}_p$ sont possibles sans permutations, alors :

— la suite de matrices $(\mathbf{A}_p)$ est convergente et :

$$\lim_{k \to +\infty} \mathbf{A}_k = \mathbf{A}_\infty \tag{3.17}$$

— $\mathbf{A}_\infty$ est une matrice triangulaire supérieure ;
— les éléments diagonaux de $\mathbf{A}_\infty$ sont les valeurs propres de $\mathbf{A}$.

Idée de la démonstration. Les matrices $\mathbf{A}_p$ sont semblables car pour tout p, $\mathbf{A}_{p+1} = \mathbf{U}_p\mathbf{L}_p$ donc $\mathbf{L}_p\mathbf{A}_{p+1} = (\mathbf{L}_p\mathbf{U}_p)\mathbf{L}_p = \mathbf{A}_p\mathbf{L}_p$ donc $\mathbf{A}_{p+1} = \mathbf{L}_p^{-1}\mathbf{A}_p\mathbf{L}_p$. Ensuite, on montre par récurrence que si $\mathbf{R}_p$ est la matrice triangulaire inférieure $\mathbf{R}_p = \mathbf{L}_0 \dots \mathbf{L}_p$ alors $\mathbf{A}_{p+1} = \mathbf{R}_p^{-1}\mathbf{A}_p\mathbf{R}_p$, puisque $\mathbf{R}_p$ admet une limite quand $p \to +\infty$ (admis). Il en résulte alors que $\mathbf{L}_p$ tend vers la matrice identité et donc que $\mathbf{A}$ tend vers une matrice triangulaire supérieure.

Remarque La convergence est de l'ordre de $\sup |\lambda_i/\lambda_j|$ où $i > j$ (admis), elle sera donc d'autant plus rapide que les valeurs propres sont espacées les unes des autres en module.

▷ **Corollaire 3.8 — Vecteurs propres de A.** Sous les hypothèses du théorème 3.7, si $\mathbf{v}$ est un vecteur propre de $\mathbf{A}_\infty$, associé à une valeur propre λ et si on note :

$$\mathbf{R}_\infty = \lim_{p\to+\infty} \mathbf{R}_p \tag{3.18}$$

$\mathbf{R}_\infty\mathbf{v}$ est un vecteur propre de $\mathbf{A}$ associé à λ.

Remarque Lorsque les hypothèses du théorème (3.7) ne sont pas vérifiées, on peut avoir une convergence vers une matrice triangulaire par blocs.

3.3.2 Matrices réelles symétriques : méthode de Jacobi

▷ **Définition 3.9 — Matrice orthogonale.** On appelle matrice orthogonale toute matrice $\mathbf{U} \in \mathbb{C}^{n,p}$ telle que :

$$\mathbf{U}^*\mathbf{U} = \mathbf{I} \quad (\text{identité dans } \mathbb{R}^{p,p}) \tag{3.19}$$

Les colonnes de $\mathbf{U}$ sont alors orthogonales.

Remarques
1. Si $\mathbf{U}$ est une matrice orthogonale, on peut avoir $\mathbf{U}^*\mathbf{U} \neq \mathbf{I}$, en particulier lorsque $p \neq n$.
2. Si $\mathbf{U}$ est une matrice réelle, carrée ($p = n$) et orthogonale alors $\mathbf{U}$ est inversible et $\mathbf{U}^{-1} = \mathbf{U}^\top$.

■ **Exemple** Soit $\boldsymbol{\omega} \in \mathbb{R}^n$ tel que $\boldsymbol{\omega}^\top\boldsymbol{\omega} = 1$ alors la matrice $\mathbf{H} = \mathbf{I} - 2\boldsymbol{\omega}\,\boldsymbol{\omega}^\top$ est une matrice orthogonale.

■ **Exemple** La matrice de rotation suivante est orthogonale :

$$\mathbf{P} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \tag{3.20}$$

▷ **Lemme 3.10** Soient $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n,n}$ une matrice symétrique et (i_0, j_0), un couple d'indices appartenant à l'ensemble $\{1, \dots, n\}^2$ avec $i_0 \neq j_0$ et soient :

$$\varphi = \frac{1}{2}\arctan\frac{2a_{i_0j_0}}{a_{j_0j_0} - a_{i_0i_0}} \tag{3.21}$$

et $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{n,n}$ définie par :

$$\begin{cases} p_{i_0i_0} = p_{j_0j_0} = \cos\varphi \\ p_{i_0j_0} = -p_{j_0i_0} = \sin\varphi \\ p_{ij} = \delta_{ij} \quad \text{sinon} \end{cases} \tag{3.22}$$

alors la matrice $\mathbf{B} = \mathbf{P}^\top \mathbf{A}\mathbf{P}$ vérifie :

$$\begin{cases} b_{ii} = a_{ii}, & \text{si } i \neq i_0 \text{ et } i \neq j_0 \\ b_{i_0 j_0} = b_{j_0 i_0} = 0 \\ b_{i_0 i_0}^2 + b_{j_0 j_0}^2 = a_{i_0 i_0}^2 + a_{j_0 j_0}^2 + 2a_{i_0 j_0} \end{cases} \tag{3.23}$$

On dit que la matrice $\mathbf{P}$ « réduit à zéro » l'élément $a_{i_0 j_0}$.

Théorème 3.11 — Méthode de Jacobi. Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice réelle symétrique. On construit par récurrence une suite de matrices $(\mathbf{B}_k)$, $\mathbf{B}_k \in \mathbb{R}^{n,n}$ par :

$$\begin{cases} \mathbf{B}_0 = \mathbf{A} \\ \mathbf{B}_k = \mathbf{P}_{k-1}^\top \mathbf{B}_{k-1} \mathbf{P}_{k-1}, \quad \forall k \in \mathbb{N}^* \end{cases} \tag{3.24}$$

où $\mathbf{P}_{k-1}$ est la matrice orthogonale $\mathbf{P}$ donnée par (3.22) avec i_0 et j_0 tels que $i_0 \neq j_0$ et :

$$|b_{i_0 j_0}^{(k-1)}| = \max_{m,\ell, m \neq \ell} |b_{m\ell}^{(k-1)}| \tag{3.25}$$

où $\mathbf{B}_k = (b_{ij}^{(k)})_{i,j=1,\dots,n}$, alors :
- la suite $(\mathbf{B}_k)$ converge vers une matrice diagonale $\mathbf{\Lambda}$ quand $k \to +\infty$;
- les éléments diagonaux de $\mathbf{\Lambda}$ sont les valeurs propres de $\mathbf{A}$.

Si de plus, on note $\mathbf{Q}_k = \mathbf{P}_0 \mathbf{P}_1 \dots \mathbf{P}_k$, alors :
- la suite $(\mathbf{Q}_k)$ a une limite notée $\mathbf{X}$ quand $k \to +\infty$;
- $\mathbf{\Lambda} = \mathbf{X}^\top \mathbf{A}\mathbf{X}$;
- les colonnes de $\mathbf{X}$ sont des vecteurs propres de $\mathbf{A}$.

Démonstration. On note la matrice $\mathbf{B}_k = (b_{ij}^{(k)})$, $i, j = 1, \dots, n$ et, d'après le lemme (3.10) :

$$b_{i_0 j_0}^{(k)} = b_{j_0 i_0}^{(k)} = 0 \tag{3.26}$$

et si $\ell \neq i_0$ et $\ell \neq j_0$:

$$b_{\ell\ell}^{(k)} = b_{\ell\ell}^{(k-1)} \tag{3.27}$$

Montrons que, sous les hypothèses du théorème (3.11) :

$$\sum_{i \neq j} (b_{ij}^{(k)})^2 \leqslant \left(1 - \frac{2}{n^2 - n}\right)^k \sum_{i \neq j} a_{ij}^2 \tag{3.28}$$

Pour tout $k \in \mathbb{N}^*$:

$$\operatorname{tr}(\mathbf{B}_k^\top \mathbf{B}_k) = \sum_{i,j=1} (b_{ij}^{(k)})^2 \tag{3.29}$$

et :

$$\sum_{i \neq j} (b_{ij}^{(k)})^2 = \operatorname{tr}(\mathbf{B}_k^\top \mathbf{B}_k) - \sum_{l=1}^{n} (b_{\ell\ell}^{(k)})^2 = \operatorname{tr}(\mathbf{B}_{k-1}^\top \mathbf{B}_{k-1}) - \sum_{l=1}^{n} (b_{\ell\ell}^{(k)})^2 \tag{3.30}$$

car la matrice $\mathbf{B}_k$ est unitairement semblable à $\mathbf{B}_{k-1}$ (c'est-à-dire qu'il existe une matrice orthogonale $\mathbf{P}$ telle que $\mathbf{B}_k = \mathbf{P}^\top \mathbf{B}_{k-1} \mathbf{P}$) donc $\mathbf{B}_k^\top \mathbf{B}$ est semblable $\mathbf{B}_{k-1}^\top \mathbf{B}_{k-1}$ et les traces sont conservées. Par conséquent, d'après le lemme 3.10 :

$$\begin{aligned} \sum_{i \neq j} (b_{ij}^{(k)})^2 &= \operatorname{tr}(\mathbf{B}_{k-1}^\top \mathbf{B}_{k-1}) - \sum_{l=1}^{n} (b_{\ell\ell}^{(k-1)})^2 - 2(b_{i_0 j_0}^{(k-1)})^2 \\ &= \sum_{i \neq j} (b_{ij}^{(k-1)})^2 - 2(b_{i_0 j_0}^{(k-1)})^2 \end{aligned} \tag{3.31}$$

Il résulte du choix de i_0 et j_0 que :

$$(b_{i_0 j_0}^{(k-1)})^2 \geqslant \Big(\frac{1}{n^2-n}\Big) \sum_{i \neq j} (b_{ij}^{(k-1)})^2 \tag{3.32}$$

d'où :

$$\sum_{i \neq j} (b_{ij}^{(k)})^2 \leqslant \Big(1 - \frac{2}{n^2-n}\Big) \sum_{i \neq j} (b_{ij}^{(k-1)})^2 \tag{3.33}$$

puis l'inégalité (3.28).

Il résulte alors de (3.28) que les termes non diagonaux de $\mathbf{B}_k$ tendent vers 0 lorsque $k \to +\infty$ et donc que la suite $(\mathbf{B}_k)$ tend vers une matrice diagonale notée $\mathbf{\Lambda}$. La suite $(\mathbf{Q}_k)$ admet donc aussi une limite, notée $\mathbf{X}$.

À la limite, nous avons $\mathbf{\Lambda} = \mathbf{X}^\top \mathbf{A}\mathbf{X}$, soit $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}$. Si $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ sont les colonnes de $\mathbf{X}$ et $\lambda_1, \lambda_2, \ldots, \lambda_n$, les éléments diagonaux de $\mathbf{\Lambda}$, alors $\lambda_1\mathbf{X}_1, \lambda_2\mathbf{X}_2, \ldots, \lambda_n\mathbf{X}_n$, sont les colonnes de $\mathbf{X}\mathbf{\Lambda}$ et par conséquent les $\lambda_1, \lambda_2, \ldots, \lambda_n$ sont les valeurs propres de $\mathbf{A}$ associées aux vecteurs propres $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$.

4

Interpolation polynomiale

▷ **Problème 4.1** Soit une fonction numérique $f : [a\,;b] \subset \mathbb{R} \to \mathbb{R}$. Soient $x_0, x_1, \ldots, x_n$, $n+1$ points distincts de $[a\,;b]$. On cherche un polynôme $p_n(x)$ de degré inférieur ou égal à n tel que :

$$p_n(x_i) = f(x_i) \quad \text{pour} \quad i = 0, \ldots, n \tag{4.1}$$

4.1 Polynôme d'interpolation de Lagrange

4.1.1 Existence et unicité du polynôme d'interpolation de Lagrange

L'existence et l'unicité de la solution sont données par le théorème classique :

Théorème 4.2 Étant donnés $f : [a\,;b] \subset \mathbb{R} \to \mathbb{R}$ et $n+1$ points distincts $x_0, x_1, \ldots, x_n$ de $[a\,;b]$ alors il existe un polynôme p_n de degré inférieur ou égal à n et un seul tel que :

$$p_n(x_i) = f(x_i) \quad \text{pour} \quad i = 0, \ldots, n \tag{4.2}$$

Ce polynôme, appelé *polynôme d'interpolation de Lagrange* de f relativement aux points $x_0, x_1, \ldots, x_n$ est donné par :

$$p_n(x) = \sum_{i=0}^{n} f(x_i) \mathrm{L}_i(x) \quad \text{avec} \quad \mathrm{L}_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k} \quad \text{pour} \quad i = 0, \ldots, n \tag{4.3}$$

Démonstration.

Unicité Si p_n et q_n sont deux polynômes répondant à la question, alors $p_n - q_n$ est un polynôme de degré inférieur ou égal à n qui admet $n+1$ racines $(x_0, x_1, \ldots, x_n)$; donc $p_n - q_n = 0$.

Existence Pour tout i, L_i est un polynôme de degré n donc $\sum_{i=0}^{n} f(x_i)\mathrm{L}_i(x)$ est un polynôme de degré inférieur ou égal à n. De plus, $\forall i, j \in \{0, \ldots, n\}$, $\mathrm{L}_i(x_j) = \delta_{ij}$. Par conséquent, le polynôme $\sum_{i=0}^{n} f(x_i)\mathrm{L}_i(x)$ est une solution.

4.1.2 Erreur d'interpolation

L'existence du polynôme d'interpolation de Lagrange ne nécessite pas de régularité particulière de la fonction f, mais elle n'est intéressante que si l'on sait mesurer l'erreur que l'on commet en remplaçant f par p_n. Or, une estimation de l'erreur d'interpolation en dehors des points d'interpolation n'est possible que si f est suffisamment régulière. En faisant des hypothèses supplémentaires sur f, on a le théorème suivant :

Théorème 4.3 — Erreur d'interpolation. Soit $\mathrm{E}(x) = f(x) - p_n(x)$, l'*erreur d'interpolation* de f en un point x de $[a\,;b]$. Si $f \in \mathcal{C}^{n+1}([a\,;b])$ alors pour tout $x \in [a\,;b]$, il existe un point ξ_x contenu dans le plus petit intervalle contenant $x_0, x_1, \ldots, x_n$ et x tel que :

$$\mathrm{E}(x) = \frac{\Pi(x)}{(n+1)!} f^{(n+1)}(\xi_x) \quad \text{où} \quad \Pi(x) = \prod_{i=0}^{n} (x - x_i) \tag{4.4}$$

Démonstration. Il faut discuter deux cas :

Cas n° 1 $x = x_j$ pour un $j \in \{0, \ldots, n\}$; alors $\mathrm{E}(x) = 0$ et $\Pi(x) = 0$.

Cas n° 2 $x \neq x_j$ pour tout $j \in \{0, \ldots, n\}$. Considérons la fonction $\mathrm{F} : [a\,;b] \to \mathbb{R}$ définie pour tout x fixé par :

$$t \to \mathrm{F}(t) = f(t) - p_n(t) - \frac{f(x) - p_n(x)}{\Pi(x)} \Pi(t) \tag{4.5}$$

alors $\mathrm{F} \in \mathcal{C}^{n+1}([a\,;b])$ et F admet $(n+2)$ zéros distincts sur $[a\,;b]$, à savoir les points x_i et x. On peut donc appliquer le théorème de Rolle entre deux zéros consécutifs de F. Il en résulte l'existence de $(n+1)$ zéros distincts pour F'. On itère le procédé n fois. Il en résulte que F'' s'annule en n points distincts de $[a\,;b]$ et, par suite, que $\mathrm{F}^{(n+1)}$ s'annule en un point de $[a\,;b]$. Soit ξ_x ce point ; nous avons donc $\mathrm{F}^{(n+1)}(\xi_x) = 0$. En dérivant $(n+1)$ fois l'expression (4.5) de F, on en déduit le résultat.

Il faut noter qu'en général ξ_x n'est pas connu. Cependant, lorsque l'on connaît une borne supérieure de $|f^{(n+1)}(x)|$, on en déduit immédiatement le résultat suivant :

▷ **Corollaire 4.4** Soit $f \in \mathcal{C}^{n+1}([a\,;b])$; on pose :

$$\mathrm{M}_{n+1} = \sup_{x \in [a;b]} \left| f^{(n+1)}(x) \right| \tag{4.6}$$

alors, une borne supérieure de l'erreur $\mathrm{E}(x)$ est donnée par :

$$|\mathrm{E}(x)| \leqslant \frac{\mathrm{M}_{n+1}}{(n+1)!} \left| \Pi(x) \right| \tag{4.7}$$

La convergence du polynôme d'interpolation vers la fonction quand n croît n'est pas vraie en général comme le montre l'exemple suivant :

■ **Exemple** Choisissons la fonction $f(x) = \exp(-x^2)$. On considère tout d'abord des points d'interpolation x_i équidistants sur l'intervalle $[-5\,;5]$. L'erreur d'interpolation peut être très grande aux extrémités de l'intervalle. C'est ce que l'on appelle les « effets de bord » de l'interpolation polynomiale, ou phénomène de Runge. En un point de cet intervalle, on peut avoir :

$$\lim_{n \to +\infty} p_n(x) \neq f(x) \tag{4.8}$$

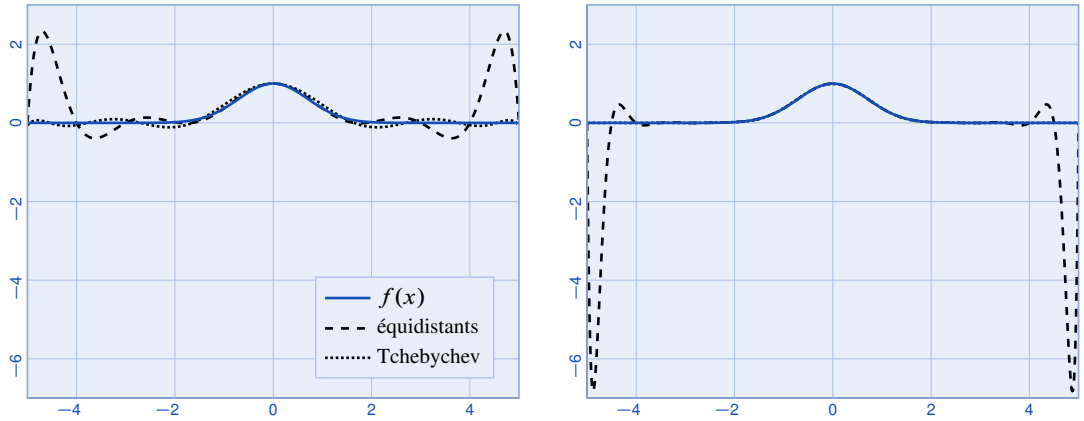
On choisit ensuite les points x_i vérifiant la relation de Tchebychev (4.10). La comparaison est illustrée sur la figure 4.1.

4.1.3 Choix des points d'interpolation

Lorsqu'il est possible de choisir les points d'interpolation (x_i), on pourra prendre les zéros du polynôme $\mathrm{P}(x)$ unitaire et de degré inférieur ou égal à $n+1$ qui minimise :

$$\inf\{\|\mathrm{P}\|_\infty ;\ \text{polynôme P-unitaire, } \deg(\mathrm{P}) \leqslant n+1\} \tag{4.9}$$

car ces points (x_i) minimisent, au sens de la norme du sup, l'erreur d'interpolation donnée par le théorème (4.3).



(a) en 11 points (b) en 21 points

Figure 4.1 – Interpolation de la fonction $f(x) = \exp(-x^2)$ sur l'intervalle $[-5\,;5]$

■ **Exemple** Lorsque $[a\,;b] = [-1\,;1]$, les points (x_i) donnant le minimum sont :

$$x_i = \cos((2i+1)\pi/(2n)), \quad i = 0, \ldots, n-1 \tag{4.10}$$

Le polynôme correspondant est le polynôme de Tchebychev :

$$\mathrm{T}_n(x) = 2^{n-1}\cos(n \arccos(x)) \tag{4.11}$$

L'erreur maximale en norme $\|\cdot\|_\infty$ est $2^{-(n-1)}$.

Sur un intervalle $[a\,;b]$, on effectue le changement de variable :

$$y = \frac{1}{2}\big((b-a)x + (a+b)\big) \tag{4.12}$$

pour se ramener à l'intervalle $[-1\,;1]$.

Nombre d'opérations
Calculer la valeur de $p_n(x)$ en un point $x = a$ à partir des formules (4.3) nécessite $n(n+1)$ multiplications, n^2 divisions et n additions.

4.2 Construction du polynôme d'interpolation

Si on ajoute un point d'interpolation x_{n+1}, les formules d'interpolation (4.3) ne permettent pas de relier directement les polynômes p_n et p_{n+1}. Cette remarque et le nombre important d'opérations dans le calcul du polynôme de Lagrange justifient la recherche d'autres constructions du polynôme d'interpolation. Les algorithmes qui suivent sont basés sur le calcul de différences de valeurs de f.

4.2.1 Différences divisées

On suppose dans ce paragraphe que les $n+1$ points x_i sont quelconques mais distincts.

▷ **Définition 4.5** Les différences divisées $f[x_i, x_{i+1}, \ldots, x_j]$ sont définies par récurrence :
— ordre 1 : $f[x_i] = f(x_i)$, $i = 0, \ldots, n$

— ordre $j-i+1$: pour $i<j$, $i,j=0,\ldots,n$

$$f[x_i,x_{i+1},\ldots,x_j]=\frac{f[x_{i+1},\ldots,x_j]-f[x_i,\ldots,x_j]}{x_j-x_i} \tag{4.13}$$

▷ **Proposition 4.6** La $(j+1)^{\text{e}}$ différence divisée $f[x_0,x_1,\ldots,x_j]$ est indépendante de l'ordre des points $x_0,x_1,\ldots,x_j$. Elle vérifie :

$$f[x_0,\ldots,x_j]=\sum_{i=0}^{j}\frac{f(x_i)}{\prod_{k=0,k\neq i}^{j}(x_i-x_k)} \tag{4.14}$$

Théorème 4.7 Le polynôme p_n qui interpole f aux points $x_0,x_1,\ldots,x_n$ s'écrit :

$$\begin{aligned}p_n(x)=\ & f[x_0]+(x-x_0)f[x_0,x_1]+(x-x_0)(x-x_1)f[x_0,x_1,x_2]+\ldots\\ &+(x-x_0)(x-x_1)\ldots(x-x_n)f[x_0,\ldots,x_n]\end{aligned} \tag{4.15}$$

L'erreur d'interpolation $\mathrm{E}(x)=f(x)-p_n(x)$ est donnée par :
— $\mathrm{E}(x)=\Pi(x)f[x_0,\ldots,x_n,x]$, $\forall x\in[a\,;b]$ tel que $x\neq x_i$ pour $i=0,\ldots,n$;
— $\mathrm{E}(x)=0$ si $x=x_i$ pour un $i\in\{0,\ldots,n\}$.

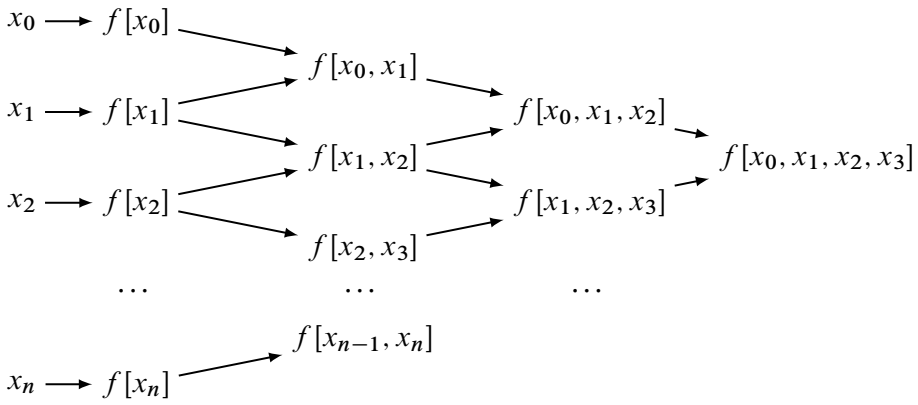
II en résulte l'algorithme de Newton de construction du polynôme P_n.

Algorithme de Newton

On construit par récurrence, une suite de polynômes p_m pour $m=0,\ldots,n$ par :

$$\begin{aligned}&p_0(x)=f[x_0]\\ &p_1(x)=p_0(x)+(x-x_0)f[x_0,x_1]\\ &\quad\ldots\ldots\ldots\\ &p_m(x)=p_{m-1}(x)+(x-x_0)(x-x_1)\ldots(x-x_{m-1})f[x_0,\ldots,x_m],\quad m=1,\ldots,n\end{aligned} \tag{4.16}$$

p_m est le polynôme de degré inférieur ou égal à m qui interpole f aux points $x_0,\ldots,x_m$. On présente ces résultats sous la forme d'un arbre :



(4.17)

Nombre d'opérations

Pour le calcul de p_n en un point x en utilisant l'algorithme de Newton, il faut $n(n+1)/2$ divisions et $n(n+1)$ additions dans (4.17) et $2n-1$ multiplications et n additions dans (4.16).

4.2.2 Différences finies

On suppose dans ce paragraphe, que les points x_i sont régulièrement espacés. Soit h le pas de cette division. Nous avons $x_i = x_0 + ih, i = 0, \ldots, n$.

▷ **Définition 4.8** L'opérateur de différences finies ∇_h est défini, si $h \neq 0$, par :

$$\nabla_h f(x) = \frac{f(x+h) - f(x)}{h} \tag{4.18}$$

et pour $m > 1$:

$$\nabla_h^m f(x) = \nabla_h^{m-1} \nabla_h \tag{4.19}$$

▷ **Lemme 4.9** Si $m \in \mathbb{N}$ alors $\nabla_h^m f(x_i) = m!\, f[x_i, x_{i+1}, \ldots, x_{i+m}]$

Ce lemme permet d'exprimer le polynôme d'interpolation de f en fonction des différences finies.

Théorème 4.10 Lorsque les points d'interpolation x_i $i = 0, \ldots, n$ sont régulièrement espacés, le polynôme p_n s'écrit :

$$\begin{aligned} p_n(x) = f(x_0) + (x - x_0)\nabla_h f(x_0) + (x - x_0)(x - x_1)\frac{\nabla_h^2 f(x_0)}{2!} + \ldots \\ + (x - x_0)(x - x_1)\ldots(x - x_{n-1})\frac{\nabla_h^n f(x_0)}{n!} \end{aligned} \tag{4.20}$$

4.3 Schéma de Hörner

Ce schéma permet de calculer la valeur d'un polynôme $\mathrm{P}(x)$ et de ses dérivées en un point a quelconque, lorsque le polynôme $\mathrm{P}(x)$ est de la forme :

$$\mathrm{P}(x) = \sum_{i=0}^{n} a_i x^{n-i}, \quad a_i \in \mathbb{R} \text{ ou } \mathbb{C} \tag{4.21}$$

Algorithme

Soit α un point de $\mathbb{R}$, alors $\mathrm{P}(\alpha)$ est donné par :

$$\begin{aligned} &b_0 = a_0 \\ &b_k = b_{k-1}\alpha + a_k, \quad k = 1, \ldots, n-1 \\ &\mathrm{P}(\alpha) = b_n = b_{n-1}\alpha + a_n \end{aligned} \tag{4.22}$$

Calcul des dérivées

La i^{e} dérivée $\mathrm{P}^{(i)}(a)$ est donnée par :

$$\begin{aligned} &b_k^{-1} = a_k, \quad k = 0, \ldots, n \\ &\text{Pour } i = 0, \ldots, n \\ &\quad b_0^i = a_0 \\ &\quad b_k^i = b_{k-1}^i \alpha + b_k^{i-1}, \quad k = 1, \ldots, n-i \\ &\quad \mathrm{P}^{(i)}(\alpha) = i!\, b_{n-i}^i \end{aligned} \tag{4.23}$$

c'est-à-dire que, pour $i = 1, \ldots, n$, les coefficients b_k^i pour $k = 0, \ldots, n-i$ sont calculés par l'algorithme de Hörner (4.22) en remplaçant les coefficients a_k par les coefficients b_k^{i-1}.

Nombre d'opérations

L'algorithme de Hörner comprend n multiplications et n additions.

Approximation de fonctions

▷ **Problème 5.1** Étant donnée une fonction f appartenant à un espace E à préciser, on cherche une fonction g appartenant à un ensemble F donné (par exemple, l'ensemble des polynômes de degré inférieur ou égal à n) telle que :

$$\forall h \in \mathrm{F}, \quad \|f - g\| \leqslant \|f - h\| \tag{5.1}$$

pour une norme $\|\cdot\|$ à préciser.

5.1 Approximation hilbertienne

Définissons un cadre fonctionnel. Soient E un espace vectoriel sur $\mathbb{R}$ (de dimension finie ou non) et F, un sous-espace vectoriel de E de dimension finie. On suppose que E est un espace préhilbertien réel, de produit scalaire $\langle\cdot,\cdot\rangle$ et de norme associée $\|\cdot\|$. Rappelons le théorème de la projection orthogonale dans le cas particulier qui nous intéresse :

Théorème 5.2 Sous les hypothèses précédentes sur E et F et pour tout $f \in \mathrm{E}$, il existe un $g \in \mathrm{F}$ et un seul tel que :

$$\forall h \in \mathrm{F}, \quad \|f - g\| \leqslant \|f - h\| \tag{5.2}$$

La condition (5.2) équivaut à :

$$\|f - g\| = \inf_{h \in \mathrm{F}} \|f - h\| \tag{5.3}$$

ou encore, puisque F est un sous-espace vectoriel de E :

$$\forall h \in \mathrm{F}, \quad \langle f - g, f - h\rangle = 0 \tag{5.4}$$

Donnons maintenant une forme explicite, et utilisable numériquement, de la solution.

Théorème 5.3 Sous les hypothèses précédentes et si de plus $(g_i)_{i=0,\ldots,n}$ est une base de F, alors la solution g de (5.2) est donnée par :

$$g = \sum_{i=0}^{n} \lambda_i g_i \tag{5.5}$$

où $\boldsymbol{\lambda} = (\lambda_i)_{i=0,\dots,n}$ est la solution du système linéaire :

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{b} \tag{5.6}$$

avec :

$$\mathbf{A} = (a_{ij})_{i,j=0,\dots,n} \qquad a_{ij} = \langle g_i, g_j \rangle, \quad i,j = 0,\dots,n \tag{5.7}$$

$$\mathbf{b} = (b_i)_{i=0,\dots,n} \qquad b_i = \langle f, g_i \rangle, \quad i = 0,\dots,n \tag{5.8}$$

Remarque Si la base $(g_i)_{i=0,\dots,n}$ est orthonormée, alors $\mathbf{A} = \mathbf{I}$ et $\boldsymbol{\lambda} = \mathbf{b}$.

Soit $\mathcal{P}_n$ l'espace vectoriel des polynômes de degré inférieur ou égal à n.

▷ **Définition 5.4** Lorsque $\mathrm{F} = \mathcal{P}_n$, une solution $g = p_n$ de (5.2) est *appelée polynôme de meilleure approximation* de f dans $(\mathrm{E}, \langle \cdot, \cdot \rangle)$ par un polynôme de degré inférieur ou égal à n.

Il est important de noter que g dépend du produit scalaire (ou de la norme) choisi sur l'espace E.

5.2 Approximation au sens des moindres carrés

5.2.1 Données dans $\mathcal{L}^2_w(a\,;b)$

La fonction f que l'on cherche à approcher est donnée sous la forme d'une fonction définie sur un intervalle de $\mathbb{R}$. Soit $(a\,;b)$ un intervalle de $\mathbb{R}$. Soit $w : x \in (a\,;b) \to w(x) \in \mathbb{R}$, une fonction poids intégrable sur $(a\,;b)$ telle que $w(x) > 0$ presque partout sur $(a\,;b)$. Soit :

$$\mathcal{L}^2_w(a\,;b) = \left\{ f : (a\,;b) \to \mathbb{R}, \int_a^b |f(x)|^2 w(x)\, \mathrm{d}x < +\infty \right\} \tag{5.9}$$

C'est un espace de Hilbert pour le produit scalaire :

$$\forall g, h \in \mathcal{L}^2_w(a\,;b), \quad \langle f, g \rangle_w = \int_a^b g(x)\, h(x)\, w(x)\, \mathrm{d}x \tag{5.10}$$

Soit F un sous-espace vectoriel de dimension finie de $\mathcal{L}^2_w(a\,;b)$. Soit $f \in \mathcal{L}^2_w(a\,;b)$; l'approximation de f dans F au sens des moindres carrés avec poids w est la meilleure approximation de f dans F pour la norme associée au produit scalaire $\langle \cdot, \cdot \rangle_w$. La solution g vérifie $g \in \mathrm{F}$ et :

$$\|f - g\|^2_w = \int_a^b (f(x) - g(x))^2\, w(x)\, \mathrm{d}x = \inf_{h \in \mathrm{F}} \int_a^b (f(x) - h(x))^2\, w(x)\, \mathrm{d}x \tag{5.11}$$

■ **Exemple** Si $(a\,;b) = [0\,;1]$, $w(x) = 1$ sur $[0\,;1]$, $\mathrm{F} = \mathcal{P}_n$ et $f \in \mathcal{L}^2_w([0\,;1])$ alors g est le polynôme de degré inférieur ou égal à n de meilleure approximation de f au sens des moindres carrés. Si la base choisie sur $\mathcal{P}_n$ est la base canonique $\{g_i : x \in (a\,;b) \to x^i,\ i = 0,\dots,n\}$ alors la matrice $\mathbf{A}$ du système linéaire (5.6) est la matrice de Hilbert, de coefficients :

$$a_{ij} = \frac{1}{i + j - 1}, \quad i,j = 0,\dots,n \tag{5.12}$$

Le mauvais conditionnement de cette matrice nécessite de faire des calculs en double précision et de ne pas choisir n trop grand.

■ **Exemple** Soit $f : \mathbb{R} \to \mathbb{R}$. Si f est 2π-périodique et de carré sommable sur l'intervalle $[-\pi\,;\pi]$, on choisit $(a\,;b) = [-\pi\,;\pi]$ et $w(x) = 1$ sur $[-\pi\,;\pi]$. L'espace F des approximations est le sous-espace vectoriel de $\mathcal{L}^2_w([-\pi\,;\pi])$ engendré par les fonctions trigonométriques $\{g_i,\ i = 0,\dots,2n+1\}$ définies par :

$$\begin{aligned} &g_0(x) = 1/2 \\ &g_{2p}(x) = \cos(2px), \qquad p = 1,\dots,n \\ &g_{2p+1}(x) = \sin((2p+1)x), \quad p = 1,\dots,n \end{aligned} \tag{5.13}$$

5.2.2 Données dans $\mathbb{R}^n$: approximation au sens des moindres carrés discret

La fonction f que l'on cherche à approcher est donnée par un nuage de points. C'est le cas, par exemple, quand la fonction f est caractérisée par un ensemble de valeurs expérimentales.

▷ **Proposition 5.5** Étant donnés $\mathrm{N}+1$ points (x_i, y_i), $i = 0,\dots,\mathrm{N}$ de $\mathbb{R}^n \times \mathbb{R}^n$, $\mathrm{N}+1$ poids positifs w_i et un entier m tel que $m \leqslant \mathrm{N}$. Alors il existe un polynôme P_m et un seul de degré inférieur ou égal à m tel que :

$$\mathrm{E}_m = \sum_{i=0}^{\mathrm{N}} (y_i - \mathrm{P}_m(x_i))^2 w_i \tag{5.14}$$

soit minimal. Ce polynôme P_m est donné par :

$$\mathrm{P}_m(x) = \sum_{k=0}^{m} a_k x^k \tag{5.15}$$

où les coefficients (a_k) sont solutions du système linéaire $\mathbf{Sa} = \mathbf{v}$ avec $\mathbf{a} = (a_k)_{k=0,\dots,m}$ et où $\mathbf{S} = (s_{kj})_{k,j=0,\dots,m}$ et $\mathbf{v} = (v_j)_{j=0,\dots,m}$ sont donnés par :

$$s_{kj} = \sum_{i=0}^{\mathrm{N}} x_i^{k+j} w_i \quad \text{et} \quad v_j = \sum_{i=0}^{\mathrm{N}} x_i^j y_i w_i, \quad k, j = 0,\dots,m \tag{5.16}$$

La proposition résulte du théorème 5.3.

5.2.3 Convergence des approximations au sens des moindres carrés

On a tout d'abord le théorème suivant :

Théorème 5.6 Soit w un poids positif intégrable sur l'intervalle fermé $[a\,;b]$. On choisit $\mathrm{F} = \mathcal{P}_n$ pour $n \in \mathbb{N}$. Soit f une fonction continue sur $[a\,;b]$ et soit $q_n \in \mathcal{P}_n$, le polynôme de degré inférieur ou égal à n qui réalise son approximation au sens des moindres carrés relativement au poids w. Alors :

$$\lim_{n\to+\infty} \|f - q_n\|_w = 0 \tag{5.17}$$

C'est une convergence en moyenne quadratique avec poids. Cette convergence n'entraîne pas la convergence uniforme. On a cependant un résultat de convergence en choisissant le poids $w = 1$ et en supposant la fonction f plus régulière.

Théorème 5.7 On choisit $\mathrm{F} = \mathcal{P}_n$ pour $n \in \mathbb{N}$. Soit f une fonction appartenant à $\mathcal{C}^2([a\,;b])$ et soit $q_n \in \mathcal{P}_n$ le polynôme de degré inférieur ou égal à n qui réalise son approximation au sens des moindres carrés relativement au poids $w = 1$. Alors :

$$\lim_{n\to+\infty} \sup_{x\in[a;b]} |f(x) - q_n(x)| = 0 \tag{5.18}$$

5.3 Polynômes orthonormés

Lorsque l'espace des approximations F est $\mathcal{P}_n$, il est souvent préférable d'utiliser une base de F qui soit orthonormée.

Théorème 5.8 Soit w une fonction poids telle que :

$$w > 0 \text{ presque partout sur }]a\,;b[\subset \mathbb{R} \tag{5.19}$$

$$\forall n \in \mathbb{N}, \quad \int_a^b w(x)x^n \,\mathrm{d}x < +\infty \tag{5.20}$$

Soit $\langle\cdot,\cdot\rangle_w$ le produit scalaire défini par :

$$\forall g, h \in \mathcal{L}^2_w(a\,;b), \quad \langle g, h\rangle_w = \int_a^b w(x)\, g(x)\, h(x) \,\mathrm{d}x < +\infty \tag{5.21}$$

alors :

1. il existe une suite de polynômes $p_0, p_1, \ldots, p_n, n \in \mathbb{N}$, vérifiant :

$$\deg(p_n) = n \quad \text{et} \quad \forall q \in \mathcal{P}_{n-1}, \quad \langle p_n, q\rangle_w = 0 \tag{5.22}$$

   Cette suite est unique à un coefficient multiplicatif près ;
2. si de plus p_n est unitaire pour tout $n \in \mathbb{N}$, la suite (p_n) vérifie les équations de récurrence :

$$p_n(x) = (x - \alpha_n)p_{n-1}(x) - \beta_n p_{n-2}(x) \tag{5.23}$$

   où :

$$\alpha_n = \frac{\langle xp_{n-1}, p_{n-1}\rangle_w}{\|p_{n-1}\|^2_w} \quad \text{et} \quad \beta_n = \frac{\|p_{n-1}\|^2_w}{\|p_{n-2}\|^2_w} \tag{5.24}$$

3. les polynômes :

$$\tilde{p}_n = \frac{p_n}{\|p_n\|_w}, \quad n \in \mathbb{N} \tag{5.25}$$

   forment une base orthonormée de l'espace vectoriel des polynômes.

Démonstration du 1. En utilisant le procédé d'orthogonalisation de Gram-Schmidt sur la base canonique $\{1, x, x^2, \ldots, x^n, \ldots\}$ de l'espace vectoriel des polynômes, on obtient :

$$\begin{aligned} &p_0(x) = 1 \\ &\text{pour } n \in \mathbb{N}^*, \quad p_n(x) = x^n - \sum_{k=0}^{n-1} \frac{\langle x^n, p_k\rangle_w}{\langle p_k, p_k\rangle_w} p_k(x) \end{aligned} \tag{5.26}$$

avec, $\forall n \in \mathbb{N}$, $\deg(p_n) = n$. Pour $n \in \mathbb{N}$, les polynômes $p_0, p_1, p_2, \ldots, p_n$ forment une base de $\mathcal{P}_n$. Ces polynômes sont orthogonaux entre eux. L'unicité résulte de l'hypothèse $\deg(p_n) = n$.

▷ **Définition 5.9** Les polynômes p_n (respectivement $\tilde{p}_n$) s'appellent les polynômes orthogonaux (respectivement orthonormés) sur l'intervalle $(a\,;b)$ pour la fonction poids w.

Une propriété utile :

▷ **Proposition 5.10** Les n racines du polynôme p_n sont réelles, distinctes et intérieures à l'intervalle $]a\,;b[$.

Théorème 5.11 Soit w un poids vérifiant les propriétés (5.19) et (5.20). Soit $f \in \mathrm{E} = \mathcal{L}^2_w([a\,;b])$, alors il existe un polynôme q_n et un seul tel que :

$$q_n \in \mathcal{P}_n \quad \text{et} \quad \|f - q_n\|_w = \inf_{q\in\mathcal{P}_n} \|f - q\|_w \tag{5.27}$$

Ce polynôme est donné par :

$$q_n(x) = \sum_{k=0}^{n} \frac{\langle f, p_k\rangle_w}{\langle p_k, p_k\rangle_w} p_k(x) \tag{5.28}$$

où les polynômes p_k vérifient (5.23) et (5.24). De plus, q_n est l'unique polynôme tel que $q_n \in \mathcal{P}_n$ et $\forall q \in \mathcal{P}_n$, $\langle q_n, q\rangle_w = \langle f, q\rangle_w$.

Quelques polynômes orthogonaux ou orthonormés sont listés dans le tableau 5.1.

$(a\,;b)$	$w(x)$	nom	polynômes
$]-1\,;1[$	1	Legendre	$p_n(x) = \dfrac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n$
$]-1\,;1[$	$\dfrac{1}{\sqrt{1-x^2}}$	Chebyshev	$\mathrm{T}_n(x) = \cos(n \arccos x)$
$]0\,;+\infty[$	$\mathrm{e}^{-\alpha x}$ $(\alpha > 0)$	Laguerre	$\mathrm{L}_n^{(\alpha)}(x) = \mathrm{e}^{\alpha x}\dfrac{\mathrm{d}^n}{\mathrm{d}x^n}(x^n\,\mathrm{e}^{-\alpha x})$
$]-\infty\,;+\infty[$	$\mathrm{e}^{-\alpha x^2}$ $(\alpha > 0)$	Hermite	$\mathrm{H}_n(x) = \mathrm{e}^{\alpha x^2}\dfrac{\mathrm{d}^n}{\mathrm{d}x^n}(\mathrm{e}^{-\alpha x^2})$

Tableau 5.1 – Polynômes orthogonaux ou orthonormés

6

Intégration numérique

▷ **Problème 6.1** Soit $f : \mathrm{I} \to \mathbb{R}$ une application continue où I est un intervalle de $\mathbb{R}$. Soit $(a\,;b) \subset \mathrm{I}$. On cherche à calculer l'intégrale :

$$\int_a^b f(x)w(x)\,\mathrm{d}x \tag{6.1}$$

pour un poids w donné, lorsque cette intégrale est définie.

6.1 Étude générale

6.1.1 Formulation

Les formules d'intégration numérique étudiées sont de la forme :

$$\int_a^b f(x)w(x)\,\mathrm{d}x = \sum_{i=0}^{k} \lambda_i f(x_i) + \mathrm{R}_{a,b}(f) \tag{6.2}$$

où :

— les points $(x_i)_{i=0,\ldots,k}$ sont $k+1$ points de l'intervalle I ;
— les $(\lambda_i)_{i=0,\ldots,k}$ sont les $k+1$ points positifs qui dépendent de la méthode ;
— $\mathrm{R}_{a,b}(f)$ est l'erreur théorique d'intégration numérique.

Ces formules peuvent être obtenues par intégration du polynôme d'interpolation p_k de f aux points $(x_i)_{i=0,\ldots,k}$. Soient en effet $x_0, \ldots, x_k$, $k+1$ points distincts de I et soit p_k le polynôme d'interpolation de f en ces $k+1$ points. Il résulte de (4.3) que :

$$p_k(x) = \sum_{i=0}^{k} f(x_i)\mathrm{L}_i(x) \tag{6.3}$$

Si w est une fonction poids positive sur l'intervalle $(a\,;b)$ telle que :

$$\forall m \in \mathbb{N}, \quad \int_a^b w(x)x^m\,\mathrm{d}x < +\infty \tag{6.4}$$

et si l'intégrale est définie, alors cette intégrale est approchée par :

$$\int_a^b p_k(x)w(x)\,\mathrm{d}x = \sum_{i=0}^{k} \lambda_i f(x_i) \quad \text{où} \quad \lambda_i = \int_a^b \mathrm{L}_i(x)w(x)\,\mathrm{d}x \tag{6.5}$$

6.1.2 Erreur d'intégration numérique

L'erreur d'intégration numérique est alors donnée par :

$$\mathrm{R}_{a,b}(f) = \int_a^b \big(f(x) - p_k(x)\big)w(x)\,\mathrm{d}x = \int_a^b \mathrm{E}(x)w(x)\,\mathrm{d}x \tag{6.6}$$

où $\mathrm{E}(x)$ désigne l'erreur d'interpolation [théorème (4.3)].

▷ **Proposition 6.2** Si $f \in \mathcal{C}^{k+1}(\mathrm{I})$ et s'il existe une constante strictement positive M_{k+1} telle que, pour tout $x \in \mathrm{I}$ on ait :

$$|f^{(k+1)}(x)| \leqslant \mathrm{M}_{k+1} \tag{6.7}$$

alors :

$$|\mathrm{R}_{a,b}(f)| \leqslant \frac{\mathrm{M}_{k+1}}{(k+1)!}\int_a^b |\Pi(x)|\,w(x)\,\mathrm{d}x \quad \text{où} \quad \Pi(x) = \prod_{i=0}^{k}(x - x_i) \tag{6.8}$$

La proposition résulte du théorème 4.3.

Afin de comparer les méthodes d'intégration numérique, on introduit une notion d'*ordre* de ces méthodes.

▷ **Définition 6.3** La formule d'intégration numérique (6.2) est dite d'ordre N si :
- $\mathrm{R}_{a,b}(p) = 0$ pour tout polynôme p de degré inférieur ou égal à N ;
- $\mathrm{R}_{a,b}(p) \neq 0$ pour un polynôme p de degré $\mathrm{N}+1$.

6.1.3 Convergence des méthodes d'intégration numérique

Les coefficients λ_i de la formule (6.2) définissant la méthode dépendent en fait du nombre k de points de base ; il conviendrait donc mieux de faire apparaître cette dépendance ; l'intégrale (6.1) est donc approchée par :

$$\mathrm{L}_k(f) = \sum_{i=0}^{k} \lambda_{ik} f(x_i) \tag{6.9}$$

▷ **Définition 6.4 — Méthode convergente.** La méthode d'intégration numérique définie par $\mathrm{L}_k(f)$ est dite *convergente* sur un ensemble V si :

$$\forall f \in \mathrm{V},\ \lim_{k\to+\infty} \mathrm{L}_k(f) = \lim_{k\to+\infty} \sum_{i=0}^{k} \lambda_{ik} f(x_i) = \int_a^b f(x)w(x)\,\mathrm{d}x \tag{6.10}$$

6.2 Formules d'intégration numérique

6.2.1 Formules élémentaires de Newton-Côtes

L'intervalle d'intégration $(a\,;b)$ est borné. Les formules de Newton-Côtes sont obtenues en choisissant les points x_i régulièrement espacés $x_i = x_0 + ih$, $i = 0, \ldots, k$ et en approchant l'intégrale :

$$\int_a^b f(x)w(x)\,\mathrm{d}x \tag{6.11}$$

par :

$$\int_a^b p_k(x)w(x)\,\mathrm{d}x \tag{6.12}$$

où p_k est le polynôme d'interpolation de f aux points x_i, $i = 0, \ldots, k$. On admettra le résultat suivant :

▷ **Proposition 6.5**

1. L'ordre d'une formule de Newton-Côtes à $k+1$ points est k si k est impair et $k+1$ si k est pair ;
2. L'erreur d'intégration est en $\mathcal{O}(h^{k+2})$ si k est impair et en $\mathcal{O}(h^{k+3})$ si k est pair.

Les formules élémentaires de Newton-Côtes ne sont utilisées que pour de petites valeurs de k et surtout pour k pair.

Les méthodes sont maintenant détaillées pour $w = 1$ sur I.

Formules de type fermé On pose $x_0 = a$ et $x_k = b$:

— $k = 1$ — méthode des trapèzes (ordre 1) :

$$\int_{x_0}^{x_1} f(x)\,\mathrm{d}x = \frac{h}{2}\big(f(a) + f(b)\big) - \frac{h^3}{12}f''(\xi), \quad \xi \in]a\,;b[, \quad h = b - a \tag{6.13}$$

— $k = 2$ — méthode de Simpson (ordre 3) :

$$\int_a^b f(x)\,\mathrm{d}x = \frac{h}{3}\Big(f(a) + 4f\Big(\frac{a+b}{2}\Big) + f(b)\Big) - \frac{h^5}{90}f^{(4)}(\xi), \quad \xi \in]a\,;b[\tag{6.14}$$

où $h = (b-a)/2$.

Formules de type ouvert On pose $x_0 = a + h\,; x_k = b - h$:

— $k = 0$ — formule du point milieu (ordre 1) :

$$\int_a^b f(x)\,\mathrm{d}x = (b-a)f\Big(\frac{a+b}{2}\Big) + \frac{h^3}{3}f''(\xi), \quad \xi \in]a\,;b[, \quad h = \frac{b-a}{2} \tag{6.15}$$

— $k = 1$ (ordre 1) :

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{2}\Big(f\Big(\frac{2a+b}{3}\Big) + f\Big(\frac{a+3b}{3}\Big)\Big) + \frac{h^3}{4}f''(\xi), \quad \xi \in]a\,;b[\tag{6.16}$$

avec $h = (b-a)/3$.

Remarque Les formules de Newton-Côtes ne sont pas convergentes ; il existe en effet des fonctions analytiques pour lesquelles :

$$\lim_{k\to+\infty}\int_a^b p_k(x)\,\mathrm{d}x \neq \int_a^b f(x)\,\mathrm{d}x \tag{6.17}$$

Par conséquent, on utilisera uniquement les formules de Newton-Côtes dans les méthodes dites *composées*.

6.2.2 Méthodes d'intégration numérique composées

L'intervalle d'intégration $[a\,;b]$ est décomposé en n intervalles $[a_i\,;a_{i+1}]$ de longueur $(b-a)/n$ tel que :

$$[a\,;b] = \bigcup_{i=0}^{n-1}[a_i\,;a_{i+1}] \tag{6.18}$$

puis, sur chacun des intervalles $[a_i\,;a_{i+1}]$, on utilise une même formule d'intégration numérique, par exemple les formules de Newton-Côtes.

Formule composée du point milieu Par définition $k = 0$ et on pose $h = (b-a)/(2n)$, $a_i = a + 2ih$ pour $i = 0, \dots, n$. En utilisant la formule du point milieu sur chacun des intervalles $[a_i\,;a_{i+1}]$, on obtient :

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \sum_{i=0}^{n-1}\int_{a_i}^{a_{i+1}} f(x)\,\mathrm{d}x \\ &= 2h\sum_{i=0}^{n-1} f\Big(\frac{a_i+a_{i+1}}{2}\Big) + \frac{b-a}{24}h^2 f''(\xi), \quad \xi\in\,]a\,;b[\end{aligned} \tag{6.19}$$

Formule composée des trapèzes Par définition $k = 1$ et on pose $h = (b-a)/n$, $a_i = a + ih$ pour $i = 0, \dots, n$. En utilisant la formule des trapèzes sur chacun des intervalles $[a_i\,;a_{i+1}]$, on obtient :

$$\int_a^b f(x)\,\mathrm{d}x = h\Big(\frac{1}{2}f(a) + \sum_{i=1}^{n-1} f(a_i) + \frac{1}{2}f(b)\Big) - \frac{b-a}{12}h^2 f''(\xi), \quad \xi\in\,]a\,;b[\tag{6.20}$$

Formule composée de Simpson Par définition $k = 2$ et on pose $h = (b-a)/(2n)$, $a_i = a + 2ih$ pour $i = 0, \dots, n$ et $x_i = a + ih$ pour $i = 0, \dots, 2n$ alors, pour $\xi \in\,]a\,;b[$:

$$\int_a^b f(x)\,\mathrm{d}x = \frac{h}{3}\sum_{i=0}^{n-1}\big(f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\big) - \frac{b-a}{180}h^4 f^{(4)}(\xi) \tag{6.21}$$

Convergence De façon générale, les méthodes d'intégration numérique composées s'écrivent, avec $w \equiv 1$ sur $[a\,;b]$:

$$\int_a^b f(x)\,\mathrm{d}x \quad \text{approché par} \quad \mathrm{F}_{nk} = \sum_{i=0}^{n-1}\Big(h\sum_{j=0}^{k}\lambda_j f(x_{ji})\Big) \tag{6.22}$$

où les coefficients λ_j sont indépendants de i.

La convergence des méthodes d'intégration numérique composées est donnée par le théorème suivant que l'on admettra :

Théorème 6.6 Si, sur chaque intervalle $[a_i\,;a_{i+1}]$, on utilise une même formule d'intégration de Newton-Côtes, pour toute fonction f intégrable sur $(a\,;b)$, on a :

$$\lim_{n\to+\infty}\mathrm{F}_{nk} = \int_a^b f(x)w(x)\,\mathrm{d}x \tag{6.23}$$

6.2.3 Formules de Gauss

L'intervalle d'intégration $(a\,;b)$ n'est plus nécessairement borné. Les points x_i et les poids λ_i, $i = 0, \dots, k$ de la formule (6.2) définissant une méthode d'intégration numérique, sont choisis ici de sorte que la méthode soit d'ordre le plus élevé possible, c'est-à-dire tels que :

$$\sum_{i=0}^{k}\lambda_i \mathrm{P}(x_i) = \int_a^b \mathrm{P}(x)w(x)\,\mathrm{d}x \tag{6.24}$$

pour tout polynôme P de degré inférieur ou égal à N, pour N le plus grand possible. On obtient le résultat voulu avec $\mathrm{N} = 2k+1$ [théorème 6.7].

Soit :

$$\langle h, g\rangle_w = \int_a^b h(x)g(x)w(x)\,\mathrm{d}x \tag{6.25}$$

un produit scalaire défini sur l'espace vectoriel $\mathcal{L}^2_w(a\,;b)$ où w est un poids vérifiant les conditions (5.19) et (5.20).

Théorème 6.7 Pour tout entier $k \in \mathbb{N}$, il existe une formule d'intégration numérique et une seule de la forme (6.2) à $k+1$ points d'ordre $2k+1$; elle est obtenue en choisissant pour x_i $(i = 0, \ldots, k)$ les racines du $(k+2)^{\text{e}}$ polynôme orthogonal p_{k+1} pour le produit scalaire (6.25) et pour poids les coefficients λ_i suivants :

$$\lambda_i = \int_a^b \mathrm{L}_i(x)w(x)\,\mathrm{d}x \quad \text{où} \quad \mathrm{L}_i(x) = \prod_{j=0, j\neq i}^{k} \frac{x - x_j}{x_i - x_j} \tag{6.26}$$

Les méthodes d'intégration numérique définies par le théorème 6.7 donnent des formules d'intégration exactes pour tout polynôme de degré inférieur ou égal $2k+1$.

Démonstration.

Unicité — Soient x_i $(i = 0, \ldots, k)$ et λ_i $(i = 0, \ldots, k)$, $k+1$ points points et $k+1$ poids définissant une méthode d'intégration numérique d'ordre $2k+1$. Soit $\Pi(x)$ le polynôme de degré $k+1$ défini par :

$$\Pi(x) = \prod_{i=0}^{k}(x - x_i) \tag{6.27}$$

Pour tout $\mathrm{Q} \in \mathcal{P}_k[x]$, $\Pi\mathrm{Q}(x) \in \mathcal{P}_{2k+1}[x]$. La formule d'intégration donnée est donc exacte pour le polynôme $\Pi\mathrm{Q}(x)$, par conséquent :

$$\int_a^b w(x)\Pi(x)\mathrm{Q}(x)\,\mathrm{d}x = \sum_{i=0}^{k} \lambda_i\,\Pi(x_i)\mathrm{Q}(x_i) = 0 \tag{6.28}$$

car $\Pi(x_i) = 0$ pour $i = 0, \ldots, n$. Le polynôme Π est donc orthogonal à tout polynôme Q de degré inférieur ou égal à k ; comme de plus le degré de Π est égal à $k+1$, Π est le $k+2$ polynôme P_{k+1}, orthogonal pour le produit scalaire $\langle\cdot,\cdot\rangle$. Il en résulte l'unicité des (x_i). On obtient les formules (6.26) en écrivant que la formule d'intégration est exacte pour les polynômes $\mathrm{L}_j(x)$.

Existence — Soient $x_0, x_1, \ldots, x_k$ les $k+1$ racines du $(k+2)^{\text{e}}$ polynôme orthogonal P_{k+1} pour le poids w sur $(a\,;b)$. Il résulte de la proposition (5.10) que ces racines sont réelles et distinctes et qu'elles appartiennent à l'intervalle $]a\,;b[$. Les fonction $\mathrm{L}_i(x)$ sont donc bien définies. Posons :

$$\lambda_i = \int_a^b w(x)\mathrm{L}_i(x)\,\mathrm{d}x, \quad i = 0, \ldots, k \tag{6.29}$$

Notons, pour tout $f \in \mathcal{C}^0(\mathrm{I})$, P_f le polynôme d'interpolation de f aux points (x_i). Nous avons :

$$\mathrm{P}_f(x) = \sum_{i=0}^{k} \mathrm{L}_i(x)\,f(x_i) \tag{6.30}$$

donc, par définition des (λ_i) :

$$\int_a^b w(x)\mathrm{P}_f(x)\,\mathrm{d}x \sum_{i=0}^{k} \lambda_i\,f(x_i) \tag{6.31}$$

— si f est un polynôme de degré inférieur ou égal à k, alors $\mathrm{P}_f = f$. Il en résulte que la formule est exacte pour les polynômes de degré inférieur ou égal à k ;

— si f est un polynôme de degré inférieur ou égal à $2k+1$, il existe des polynômes r et q tels que :

$$f(x) = q(x)\prod_{i=0}^{k}(x - x_i) + r(x) \quad \text{avec } \deg r \leqslant k \text{ et } \deg q \leqslant k \tag{6.32}$$

par conséquent :

$$\int_a^b w(x)\,f(x)\,\mathrm{d}x = \int_a^b w(x)r(x)\,\mathrm{d}x = \sum_{i=0}^{k} \lambda_i\, r(x_i) = \sum_{i=0}^{k} \lambda_i\, f(x_i) \tag{6.33}$$

car $\mathrm{P}_{k+1} = \prod(x - x_i)$ est orthogonal à tout polynôme de degré inférieur ou égal à k, et parce que la formule d'intégration numérique est exacte sur les polynômes de degré inférieur ou égal à k (donc pour le polynôme r). La formule est donc aussi exacte pour tout polynôme de degré inférieur ou égal à $2k + 1$.

▷ **Définition 6.8** Les formules d'intégration numérique données par le théorème 6.7 sont appelées les formules de Gauss et en particulier : formules de Gauss-Legendre (resp. de Gauss-Chebyshev...) si les polynômes utilisés sont les polynômes de Legendre (resp. de Chebyshev...).

Remarque Les coefficients λ_i sont aussi donnés par :

$$\lambda_i = \frac{1}{p'_{k+1}(x_i)} \int_a^b \frac{p_{k+1}(x)}{x - x_i} w(x)\,\mathrm{d}x \tag{6.34}$$

Ils sont tabulés pour un certain nombre de polynômes classiques.

■ **Exemple — Formules de Gauss-Legendre.** L'intervalle est $(a\,;b) = [-1\,;1]$ avec la fonction de pondération $w = 1$. Les polynômes utilisés sont les polynômes de Legendre. Le nombre de points de base utilisé est $n = k + 1$.

k	n	ordre	λ_i	x_i
0	1	1	$\lambda_0 = 2$	$x_0 = 0$
1	2	3	$\lambda_0 = \lambda_1 = 1$	$x_1 = -x_0 = 1/\sqrt{3}$
2	3	5	$\lambda_0 = \lambda_2 = 5/9$	$x_2 = -x_0 = 0{,}77459667$
			$\lambda_1 = 8/9$	$x_1 = 0$
3	4	7	$\lambda_0 = \lambda_3 = 0{,}34785484$	$x_3 = -x_0 = 0{,}86113631$
			$\lambda_1 = \lambda_2 = 0{,}65214515$	$x_2 = -x_1 = 0{,}33998104$
4	5	9	$\lambda_0 = \lambda_4 = 0{,}23692688$	$x_4 = -x_0 = 0{,}90617984$
			$\lambda_1 = \lambda_3 = 0{,}47862867$	$x_3 = -x_1 = 0{,}53846931$
			$\lambda_2 = 0{,}56888888$	$x_2 = 0$
5	6	11	$\lambda_0 = \lambda_5 = 0{,}17132449$	$x_5 = -x_0 = 0{,}93246951$
			$\lambda_1 = \lambda_4 = 0{,}36076157$	$x_4 = -x_1 = 0{,}66120930$
			$\lambda_2 = \lambda_3 = 0{,}46791393$	$x_3 = -x_2 = 0{,}23861919$

Tableau 6.1 – Caractérisation de l'intégration avec les polynômes de Legendre

■ **Exemple — Formules de Gauss-Hermite.** L'intervalle est $(a\,;b) =]-\infty\,;\infty[$ avec la fonction de pondération $w = \exp(-x^2)$. Les polynômes utilisés sont les polynômes d'Hermite. Le nombre de points est $n = k + 1$.

6.3 Intégration numérique en deux dimensions

Les résultats du théorème 6.7 concernant les formules d'intégration numérique de Gauss peuvent s'étendre à l'intégration de certaines fonctions sur des domaines de $\mathbb{R}^2$ ou $\mathbb{R}^3$. Nous ne développerons pas ici la théorie en deux et trois dimensions des formules d'intégration de Gauss mais donnons quelques exemples de formules très utilisées en analyse numérique, notamment pour le calcul d'intégrales apparaissant dans la méthode des éléments finis.

k	n	ordre	λ_i	x_i
0	1	1	$\lambda_0 = 1{,}77245385$	$x_0 = 0$
1	2	3	$\lambda_0 = \lambda_1 = 0{,}88622693$	$x_1 = -x_0 = 0{,}70710678$
2	3	5	$\lambda_0 = \lambda_2 = 0{,}29540898$	$x_2 = -x_0 = 1{,}22474487$
			$\lambda_1 = 1{,}1815359$	$x_1 = 0$
3	4	7	$\lambda_0 = \lambda_3 = 0{,}08131283$	$x_3 = -x_0 = 1{,}65068012$
			$\lambda_1 = \lambda_2 = 0{,}80491409$	$x_2 = -x_1 = 0{,}52464762$
4	5	9	$\lambda_0 = \lambda_4 = 0{,}01995324$	$x_4 = -x_0 = 2{,}02018287$
			$\lambda_1 = \lambda_3 = 0{,}39361932$	$x_3 = -x_1 = 0{,}95857246$
			$\lambda_2 = 0{,}94530872$	$x_2 = 0$
5	6	11	$\lambda_0 = \lambda_5 = 0{,}00453001$	$x_5 = -x_0 = 2{,}35060497$
			$\lambda_1 = \lambda_4 = 0{,}15706732$	$x_4 = -x_1 = 1{,}33584907$
			$\lambda_2 = \lambda_3 = 0{,}72462960$	$x_3 = -x_2 = 0{,}43607741$

Tableau 6.2 – Caractérisation de l'intégration avec les polynômes d'Hermite

■ **Exemple — Intégration sur un carré de $\mathbb{R}^2$.** Soit D le carré de $\mathbb{R}^2$ défini par $\mathrm{D} = [-1\,;1] \times [-1\,;1]$. Si $f \in \mathcal{C}^0([-1\,;1] \times [-1\,;1]; \mathbb{R})$, on approche l'intégrale de f sur le carré D par une formule de la forme :

$$\int_{-1}^{1}\int_{-1}^{1} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \mathrm{I}_k + \mathrm{R}_{\mathrm{D}(f)} \tag{6.35}$$

où :

$$\mathrm{I}_k = \sum_{i=0}^{k}\sum_{j=0}^{k} \lambda_i \lambda_j f(x_i, y_j) \tag{6.36}$$

pour un entier k donné, où les λ_i sont des coefficients réels, et les (x_i, y_j) sont des points du domaine :

Formule à un point de Gauss-Legendre On applique la formule (6.36) pour $k = 0$:

$$\mathrm{I}_0 = 4f(0,0) \tag{6.37}$$

Cette formule est exacte sur $\mathrm{Q}_1 = \mathrm{Vect}(1, x, y, xy)$.

Formule à quatre points de Gauss-Legendre On applique la formule (6.36) pour $k = 1$:

$$\mathrm{I}_1 = \sum_{i=0}^{1}\sum_{j=0}^{1} f(x_i, y_j) \quad \text{avec} \quad (x_i, y_j) = (\pm 1/\sqrt{3}, \pm 1/\sqrt{3}) \tag{6.38}$$

Cette formule est exacte sur $\mathrm{Q}_3 = \mathrm{Vect}(x^i y^j,\ 0 \leqslant i, j \leqslant 3)$.

■ **Exemple — Intégration sur un triangle de $\mathbb{R}^2$.** Soit T un triangle rectangle isocèle défini par $\mathrm{T} = \{(x,y);\ 0 \leqslant y \leqslant 1,\ 0 \leqslant x \leqslant 1-y\}$. Si $f \in \mathcal{C}^0(\mathrm{T}; \mathbb{R})$, on approche l'intégrale de f sur le triangle T par une formule de la forme :

$$\int_{0}^{1}\int_{0}^{1-y} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \mathrm{I}_k + \mathrm{R}_{\mathrm{T}(f)} \tag{6.39}$$

où :

$$\mathrm{I}_k = \sum_{i=0}^{k} \lambda_i f(x_i, y_i) \tag{6.40}$$

pour un entier k donné et où les λ_i sont k coefficients réels, et les (x_i, y_i), k points du domaine :

Formule à un point de Gauss-Legendre On applique la formule (6.40) pour $k = 0$:

$$\mathrm{I}_0 = \frac{1}{2} f(1/3, 1/3) \tag{6.41}$$

Cette formule est exacte sur $\mathrm{P}_1 = \mathrm{Vect}(1, x, y)$.

Formule à trois points de Gauss-Legendre On applique la formule (6.40) pour $k = 2$:

$$\mathrm{I}_2 = \frac{1}{6} f(1/6, 1/6) + \frac{1}{6} f(2/3, 1/6) + \frac{1}{6} f(1/6, 2/3) \tag{6.42}$$

Cette formule est exacte sur $\mathrm{P}_3 = \mathrm{Vect}(x^i y^j,\ 0 \leqslant i + j \leqslant 3)$.

7

Équations différentielles du premier ordre à condition initiale

7.1 Problème de Cauchy

7.1.1 Condition de Cauchy

Soit $[x_0\,;x_0+a]$ un intervalle de $\mathbb{R}$ où x_0 et $a>0$ sont donnés. Soit $\mathbf{f}:[x_0\,;x_0+a]\times\mathbb{R}^m\to\mathbb{R}^m$ une application continue sur $[x_0\,;x_0+a]\times\mathbb{R}^m$. On cherche une fonction $\mathbf{y}$ continûment dérivable sur $[x_0\,;x_0+a]\to\mathbb{R}^m$ vérifiant :

$$\mathbf{y}'(x)=\mathbf{f}(x,\mathbf{y}(x)),\quad \forall x\in[x_0\,;x_0+a] \tag{7.1}$$

Une solution $\mathbf{y}$ de (7.1) est appelée *intégrale* de l'équation différentielle. L'équation (7.1) est équivalente à un système de m équations différentielles :

$$\begin{cases} y_1'(x)=f_1\big(x,y_1(x),y_2(x),\ldots,y_m(x)\big)\\ y_2'(x)=f_2\big(x,y_1(x),y_2(x),\ldots,y_m(x)\big)\\ \quad\cdots\cdots\\ y_m'(x)=f_m\big(x,y_1(x),y_2(x),\ldots,y_m(x)\big)\end{cases} \tag{7.2}$$

où $\mathbf{y}(x)=\big(y_1(x)\ldots y_m(x)\big)^\top$ et $\mathbf{f}=(f_1\ldots f_m)^\top$.

Remarque Une équation différentielle d'ordre p de la forme suivante $y^{(p)}=f\big(x,y,y',\ldots,y^{(p-1)}\big)$ où $p>1$ se ramène à un système équivalent de la forme (7.1) en posant :

$$\begin{cases} z_1'(x)=z_2(x)\\ z_2'(x)=z_3(x)\\ \quad\cdots\cdots\\ z_p'(x)=f\big(x,z_1(x),z_2(x),\ldots,z_p(x)\big)\end{cases} \tag{7.3}$$

autrement dit :

$$\mathbf{z}'=\mathbf{F}(x,\mathbf{z}) \tag{7.4}$$

où $\mathbf{z}=(z_1,z_2,\ldots,z_p)^\top$ et $\mathbf{z}\in\mathcal{C}^1\big([x_0\,;x_0+a];\mathbb{R}^p\big)$.

Condition de Cauchy

On se donne $\mathbf{y}_0\in\mathbb{R}^m$; on cherche $\mathbf{y}$ vérifiant (7.1) et la condition initiale :

$$\mathbf{y}(x_0)=\mathbf{y}_0 \tag{7.5}$$

Cette condition est appelée *condition de Cauchy*.

7.1.2 Théorème d'existence et d'unicité

Soit $\|\cdot\|$ une norme sur $\mathbb{R}^m$. On rappelle le théorème d'existence et d'unicité classique suivant qui nous donne le cadre mathématique dans lequel nous chercherons une méthode de résolution numérique des équations (7.1) et (7.5).

Théorème 7.1 Soit $\mathbf{f}$ une application continue :

$$\mathbf{f} : (x, \mathbf{z}) \in [x_0 ; x_0 + a] \times \mathbb{R}^m \to \mathbf{f}(x, \mathbf{z}) \in \mathbb{R}^m \tag{7.6}$$

Si $\mathbf{f}$ vérifie la condition de Lipschitz :

$$\exists \mathrm{L} > 0,\ \forall x \in [x_0 ; x_0 + a],\ \forall \mathbf{z}_1 \text{ et } \mathbf{z}_2 \in \mathbb{R}^m, \quad \|\mathbf{f}(x, \mathbf{z}_1) - \mathbf{f}(x, \mathbf{z}_2)\| \leqslant \mathrm{L}\|\mathbf{z}_1 - \mathbf{z}_2\| \tag{7.7}$$

alors, il existe une fonction $\mathbf{y}$ unique appartenant à $\mathcal{C}^1([x_0 ; x_0 + a]; \mathbb{R}^m)$ et vérifiant :

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)) \qquad \forall x \in [x_0 ; x_0 + a] \tag{7.8}$$

$$\mathbf{y}(x_0) = \mathbf{y}_0 \qquad \mathbf{y}_0 \text{ donné dans } \mathbb{R}^m \tag{7.9}$$

7.2 Méthodes de résolution numériques à un pas

On se place sous les hypothèses du théorème 7.1. La condition initiale (7.9) n'étant pas toujours calculée exactement, dans la suite on notera $\boldsymbol{\eta}$ la valeur de cette condition ; on cherche donc une approximation de la solution $\mathbf{y}$ du problème suivant :

$$\begin{aligned} &\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \quad \forall x \in [x_0 ; x_0 + a] \\ &\mathbf{y}(x_0) = \boldsymbol{\eta} \end{aligned} \tag{7.10}$$

On pose, pour $n \in \mathbb{N}$ donné, $x_i = x_0 + ih$, $i = 0, \ldots, n$ avec $nh = a$. Les points $(x_i)_{i=0,\ldots,n}$ définissent une subdivision régulière de l'intervalle d'étude $[x_0 ; x_0 + a]$ et h est appelé le *pas de la discrétisation*. On cherche une approximation de $\mathbf{y}$ aux points $(x_i)_{i=0,\ldots,n}$.

7.2.1 Méthode d'Euler-Cauchy

On construit par récurrence une suite $(\mathbf{y}_i)_{i=0,\ldots,n}$ par :

$$\begin{aligned} &\mathbf{y}_0 = \boldsymbol{\eta} \\ &\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{f}(x_i, \mathbf{y}_i), \quad i = 0, \ldots, n-1 \end{aligned} \tag{7.11}$$

où $\mathbf{y}_i$ est une approximation de $\mathbf{y}(x_i)$. La dérivée $\mathbf{y}'(x_i)$ est approchée par le rapport :

$$\mathbf{y}'(x_i) \approx \frac{\mathbf{y}_{i+1} - \mathbf{y}_i}{h} \tag{7.12}$$

La quantité $\mathbf{e}_i = \mathbf{y}(x_i) - \mathbf{y}_i$ est l'erreur de discrétisation au point x_i. On démontre le théorème d'estimation d'erreur suivant :

Théorème 7.2 Soit $\mathbf{f}$ une application de $[x_0 ; x_0 + a] \times \mathbb{R}^m$ dans $\mathbb{R}^m$ vérifiant la condition de Lipschitz (7.7) ; si de plus, $\mathbf{f}$ est continûment différentiable dans :

$$\mathrm{D} = \left\{(x, \mathbf{y});\ x \in [x_0 ; x_0 + a],\ \|\mathbf{y}\| \leqslant \|\mathbf{y}_0\|\, \mathrm{e}^{\mathrm{L}a} + \frac{1}{\mathrm{L}}(\mathrm{e}^{\mathrm{L}a} - 1)\right\} \tag{7.13}$$

alors la solution $\mathbf{y}$ de (7.10) appartient à $\mathcal{C}^2([x_0\,;x_0+a])$ et l'erreur de discrétisation vérifie :

$$\|\mathbf{e}_i\| \leqslant \left(\|\mathbf{e}_0\| + \frac{h\mathrm{M}}{2\mathrm{L}}\right) \mathrm{e}^{\mathrm{L}(x_i-x_0)} \quad \text{où} \quad \mathrm{M} = \sup_{x\in[x_0;x_0+a]} \|\mathbf{y}''(x)\| \tag{7.14}$$

7.2.2 Étude générale des méthodes à un pas

Dans les méthodes à un pas, $\mathbf{y}_{i+1}$ est calculé à partir de x_i, $\mathbf{y}_i$ et h, et éventuellement à partir de points intermédiaires. Dans ce paragraphe et les suivants, on supposera que $\mathbb{R}^m = \mathbb{R}$ pour simplifier les écritures mais ces méthodes s'utilisent aussi dans $\mathbb{R}^m$. La forme générale est la suivante :

$$\begin{cases} y_0 = \eta_h \\ y_{i+1} = y_i + h\Phi(x_i, y_i, h), \quad i = 0, \ldots, n-1 \end{cases} \tag{7.15}$$

La méthode est déterminée par le choix de la fonction Φ.

▷ **Définition 7.3 — Convergence.** La méthode (7.15) est dite convergente si :

$$\lim_{n\to+\infty} \max_{i=0,\ldots,n} |y_i - y(x_i)| = 0 \quad \text{lorsque} \quad \lim_{h\to 0} \eta_h = \eta \tag{7.16}$$

▷ **Définition 7.4 — Stabilité théorique.** Soient $\{y_i, i = 0, \ldots, n\}$ et $\{z_i, i = 0, \ldots, n\}$ les solutions respectives des systèmes :

$$\begin{cases} y_0 \text{ fixé} \\ y_{i+1} = y_i + h\Phi(x_i, y_i, h), \quad i = 0, \ldots, n-1 \end{cases} \tag{7.17}$$

et :

$$\begin{cases} z_0 \text{ fixé} \\ z_{i+1} = z_i + h\Phi(x_i, z_i, h) + \varepsilon_i, \quad i = 0, \ldots, n-1 \end{cases} \tag{7.18}$$

La méthode (7.15) est dite *stable* s'il existe deux constantes M_1 et M_2 indépendantes de h telles que :

$$\max_{i=0,\ldots,n} |y_i - z_i| \leqslant \mathrm{M}_1 |y_0 - z_0| + \mathrm{M}_2 \sum_{i=0}^{n-1} |\varepsilon_i| \tag{7.19}$$

Cette notion de stabilité signifie qu'une petite perturbation sur les données n'entraîne qu'une perturbation contrôlée sur la solution calculée.

Théorème 7.5 Si la fonction Φ définissant la méthode, vérifie une condition de Lipschitz en y pour h assez petit : $\exists \Lambda > 0$ tel que $\forall x \in [x_0\,;x_0+a]$, $\forall y$ et $z \in \mathbb{R}$, $\forall h \in]0\,;h_0]$,

$$|\Phi(x, y, h) - \Phi(x, z, h)| \leqslant \Lambda |y - z| \tag{7.20}$$

alors la méthode (7.15) est stable. De plus, avec les notations de la définition 7.3 :

$$\max_{i=0,\ldots,n} |y_i - z_i| \leqslant \mathrm{e}^{a\Lambda} |y_0 - z_0| + \mathrm{e}^{a\Lambda} \sum_{i=0}^{n-1} |\varepsilon_i| \tag{7.21}$$

Pour montrer ce résultat, on utilise le Lemme de Gronvall :

▷ **Lemme 7.6 — Lemme de Gronvall.** Soit (θ_i) et (α_i) deux suites de réels positifs ou nuls vérifiant :

$$\forall i \geqslant 0, \quad \theta_{i+1} \leqslant (1 + h\mathrm{L})\theta_i + \alpha_i \tag{7.22}$$

alors :

$$\forall i \geqslant 0, \quad \theta_{i+1} \leqslant \mathrm{e}^{\mathrm{L}ih}\,\theta_0 + \sum_{j=0}^{i-1} \mathrm{e}^{\mathrm{L}(i-j-1)h}\,\alpha_j \tag{7.23}$$

Voici deux notions voisines pour mesurer la qualité du schéma d'approximation :

▷ **Définition 7.7 — Consistance.** La méthode (7.15) est dite consistante avec l'équation (7.1) si :

$$\lim_{n\to+\infty} \sum_{i=0}^{n-1} \left| y(x_{i+1} - y(x_i) - h\Phi(x_i, y(x_i), h) \right| = 0 \tag{7.24}$$

où y est une solution quelconque de l'équation (7.1).

▷ **Définition 7.8 — Ordre d'une méthode.** La méthode (7.15) est dite d'ordre p si :

$$\sum_{i=0}^{n-1} |y(x_{i+1}) - y(x_i) - h\Phi(x_i, y(x_i), h)| \leqslant \mathrm{K}\,h^p \tag{7.25}$$

où y une solution quelconque de I'équation (7.1) et K est une constante non nulle indépendante de h (mais dépendant de y et de Φ).

Théorème 7.9 Si la méthode (7.15) est stable et consistante, alors elle est convergente.

Théorème 7.10 Si la fonction Φ vérifie la condition de Lipschitz (7.20) et si la méthode (7.15) est d'ordre p alors :

$$\forall i = 1, \ldots, n, \quad |y_i - y(x_i)| \leqslant \mathrm{e}^{\Lambda(x_i - x_0)} \left(|\eta_h - \eta| + \frac{\mathrm{K}\,h^p}{\mathrm{L}} \right) \tag{7.26}$$

Remarque La méthode numérique (7.15) ne donne des approximations de la solution y qu'aux $(n + 1)$ points x_i. On peut définir une solution approchée sur tout l'intervalle $[x_0\,;x_0 + a]$ par interpolation. Par exemple, par une interpolation linéaire :

$$y_h(x) = y_i + \frac{y_{i+1} - y_i}{h}(x - x_i) \quad \text{si} \quad x_i \leqslant x \leqslant x_{i+1}, \quad i = 0, \ldots, n-1 \tag{7.27}$$

Théorème 7.11 Sous les hypothèses des théorèmes (7.5) à (7.21) :

$$\lim_{h\to 0} y_h = y \quad \text{dans} \quad \mathcal{C}^0([x_0\,;x_0 + a]) \tag{7.28}$$

■ **Exemple — Méthode d'Euler.** Elle est donnée par :

$$\Phi(x, y, h) = f(x, y) \tag{7.29}$$

C'est une méthode d'ordre 1. Elle est stable dès que f est lipschitzienne en y.

■ **Exemple — Méthodes de Runge-Kutta à q points intermédiaires.** Soient q réels $c_1, c_2, \ldots, c_q$ distincts ou non définissant q points intermédiaires $x_{ik} = x_i + c_k h$, $k = 1, \ldots, q$, $i = 0, \ldots, n$. Les méthodes de Runge-Kutta consistent à calculer successivement :

$$y_{ij} = y_i + h \sum_{k=1}^{q} a_{jk} f(x_{ik}, y_{ik}), \quad j = 1, \ldots, q \tag{7.30}$$

$$y_{i+1} = y_i + h \sum_{k=1}^{q} b_k f(x_{ik}, y_{ik}), \quad j = 1, \ldots, q \tag{7.31}$$

Ces équations peuvent être explicites ou implicites :

— la méthode est explicite quand les termes y_{ij} sont connus pour tout j ; la matrice $\mathbf{A} = (a_{ij})$ est alors triangulaire inférieure avec $a_{ij} = 0$ si $j \geqslant i$;

— la méthode est implicite sinon ; on a alors :

$$y_{ij} = h a_{jj} f(x_{ij}, y_{ij}) + h \sum_{k>j} a_{jk} f(x_{ik}, y_{ik}) + \text{termes connus} \tag{7.32}$$

Ce schéma s'écrit sous la forme générale :

$$y_{i+1} = y_i + h\Phi(x_i, y_i, h) \tag{7.33}$$

où la fonction Φ est définie par :

$$\begin{cases} \Phi(x_i, y_i, h) = \displaystyle\sum_{k=1}^{q} b_k f(x_{ik}, y_{ik}) \\ y_{ij} = y_i + h \displaystyle\sum_{k=1}^{q} a_{jk} f(x_{ik}, y_{ik}), \quad j = 1, \ldots, q \end{cases} \tag{7.34}$$

et $x_{ik} = x_i + c_k h$ pour $k = 1, \ldots, q$.

■ **Exemple — Méthode d'ordre 1.** Par définition, on a $q = 1$ et le schéma s'écrit :

$$\begin{aligned} y_{i1} &= y_i \\ y_{i+1} &= y_i + h f(x_i, y_i) \end{aligned} \tag{7.35}$$

On retrouve la méthode d'Euler.

■ **Exemple — Méthode d'ordre 2.** Par définition, on a $q = 2$. Soit un paramètre $\alpha \neq 0$. Les méthodes de Runge-Kutta définies par :

$$y_{i+1} = y_i + h\Big(1 - \frac{1}{2\alpha}\Big) f(x_i, y_i) + \frac{h}{2\alpha} f(x_i + \alpha h, y_i + \alpha h f(x_i, y_i)) \tag{7.36}$$

sont des méthodes d'ordre 2 :

— lorsque $\alpha = 1/2$, c'est la méthode d'Euler modifiée ;

— quand $\alpha = 1$, c'est la méthode de Heun.

■ **Exemple — Méthode explicite d'ordre 4.** Par définition, on a $q = 4$. Une méthode explicite classique d'ordre 4 est donnée par :

$$y_{i+1} = y_i + h\Big(\frac{1}{6} f(x_i, y_i) + \frac{1}{3} f(x_i + h/2, y_{i2}) + \frac{1}{3} f(x_i + h/2, y_{i3}) + \frac{1}{6} f(x_{i+1}, y_{i4})\Big) \tag{7.37}$$

avec :

$$\begin{aligned} y_{i1} &= y_i \\ y_{i2} &= y_i + h/2 f(x_i, y_{i1}) \\ y_{i3} &= y_i + h/2 f(x_i + h/2, y_{i2}) \\ y_{i4} &= y_i + h f(x_i + h/2, y_{i3}) \end{aligned} \tag{7.38}$$

Cette méthode s'écrit aussi plus couramment :

$$\begin{aligned} k_1 &= f(x_i, y_i) \\ k_2 &= f(x_i + h/2, y_i + h/2k_1) \\ k_3 &= f(x_i + h/2, y_i + h/2k_2) \\ k_4 &= f(x_i + h, y_i + hk_3) \\ y_{i+1} &= y_i + (k_1 + 2k_2 + 2k_3 + k_4)/6 \end{aligned} \tag{7.39}$$

7.3 Méthodes de résolution numérique à pas multiples

Les méthodes de Runge-Kutta présentent l'inconvénient de faire intervenir le calcul de $f(x, y)$ en des points ne servant pas directement au calcul. Les méthodes à pas multiples n'ont pas cet inconvénient.

On cherche toujours à résoudre numériquement le problème (7.10) où f vérifie les hypothèses de continuité et la condition de Lipschitz du théorème 7.1. On définit, comme précédemment, une subdivision régulière $(x_i)_{i=0,\dots,n}$ de l'intervalle $[x_0 ; x_0 + a]$ et on note encore y_i une approximation de $y(x_i)$. Dans une méthode à k pas, y_{i+1} est calculé à partir des k valeurs précédentes $y_i, y_{i-1}, \dots, y_{i-k+1}$ sans faire intervenir de valeurs intermédiaires.

7.3.1 Méthodes d'Adams-Bashforth à $k+1$ pas

Elles sont obtenues en intégrant sur l'intervalle $[x_i ; x_{i+1}]$ le polynôme p d'interpolation de f de degré inférieur ou égal à k aux points $x_i, x_{i-1}, \dots, x_{i-k}$. On obtient un schéma de la forme suivante :

$$y_{i+1} - y_i = \int_{x_i}^{x_{i+1}} p(\xi)\, \mathrm{d}\xi = h \sum_{\ell=0}^{k} \beta_{\ell k} f_{i-\ell} \tag{7.40}$$

où $f_i = f(x_i, y_i)$. Les coefficients $\beta_{\ell k}$ sont donnés par :

$$\beta_{\ell k} = \int_0^1 \prod_{j=0, j\neq\ell}^{k} \frac{s+j}{-\ell+j}\, \mathrm{d}s \tag{7.41}$$

Ces valeurs sont tabulées pour différentes valeurs des constantes k et ℓ. Les méthodes d'Adams-Bashforth sont explicites.

Exemples

— $k = 0$: il s'agit de la méthode d'Euler qui s'écrit $y_{i+1} = y_i + h f_i$;
— $k = 1$: $y_{i+1} = y_i + h(3f_i - f_{i-1})/2$;
— $k = 2$: $y_{i+1} = y_i + h(23f_i - 16f_{i-1} + 5f_{i-2})/12$;

Le terme y_0 est donné mais ces formules ne permettent pas de calculer les premiers termes $y_1, y_2, \dots, y_{k-1}$: la méthode n'est pas *auto-démarrante* ; le démarrage est assuré par une méthode à un pas, par exemple une méthode de Runge-Kutta.

Les méthodes d'Adams-Bashforth ne sont pas utilisées seules car elles sont en général instables.

7.3.2 Méthodes d'Adams-Moulton

Forme générale

Elles sont obtenues en intégrant, sur l'intervalle $[x_i\,;x_{i+1}]$, le polynôme d'interpolation, de degré inférieur ou égal à k, de f aux points $x_{i+1}, x_i, \ldots, x_{i-k+1}$. Elles sont de la forme suivante :

$$y_{i+1} - y_i = h \sum_{\ell=-1}^{k-1} \beta^*_{\ell k} f_{i-\ell} \tag{7.42}$$

où $f_i = f(x_i, y_i)$. Les coefficients sont donnés par :

$$\beta^*_{\ell k} = \int_{-1}^{0} \prod_{j=0, j\neq\ell}^{k} \frac{s+j}{-\ell+j}\,\mathrm{d}s \tag{7.43}$$

La relation (7.42) s'écrit encore :

$$y_{i+1} - h\beta^*_{-1\,k} f(x_{i+1}, y_{i+1}) = y_i + h \sum_{\ell=0}^{k-1} \beta^*_{\ell k} f_{i-\ell} \quad \text{avec} \quad \beta^*_{-1\,k} \neq 0 \tag{7.44}$$

Ces méthodes sont implicites.

■ **Exemple — Méthode d'Euler rétrograde.** On a $k = 0$. C'est une méthode d'ordre 1 qui s'écrit :

$$y_{i+1} = y_i + h f_{i+1} \tag{7.45}$$

■ **Exemple — Méthode de Crank-Nicolson.** On a $k = 1$. C'est une méthode d'ordre 2 qui s'écrit :

$$y_{i+1} = y_i + h\Big(\frac{f_i + f_{i+1}}{2}\Big) \tag{7.46}$$

■ **Exemple** Pour $k = 2$, la méthode est d'ordre 3 et s'écrit :

$$y_{i+1} = y_i + \frac{h}{12}(5f_{i+1} + 2f_i - f_{i-1}) \tag{7.47}$$

Résolution numérique : méthodes de prédiction-correction

Les méthodes d'Adams-Moulton sont implicites ; pour calculer y_{i+i}, il faut résoudre une équation non linéaire de la forme :

$$y_{i+1} = \mathrm{C}hf(x_{i+1}, y_{i+1}) + \mathrm{A}_i \tag{7.48}$$

où C est une constante indépendante de h et de y_{i+1} et où A_i est une quantité constituée de termes connus à l'étape $i+1$. On résout généralement cette équation par la méthode des approximations successives.

La méthode est la suivante : étant donnée $y^{(0)}_{i+1}$ une approximation initiale de y_{i+1}, on construit une suite $y^{(j)}_{i+1}$ par récurrence à partir de $y^{(0)}_{i+1}$ par :

$$y^{(j)}_{i+1} = \mathrm{C}hf(x_{i+1}, y^{(j-1)}_{i+1}) + \mathrm{A}_i, \quad j \in \mathbb{N}^* \tag{7.49}$$

On vérifie que pour h assez petit, la suite $y^{(j)}_{i+1}$ est convergente lorsque $j \to \infty$ et que la limite est l'unique solution y_{i+1} de (7.48). On montre que la méthode est (théoriquement) convergente lorsque $0 < h < 1/(c\mathrm{L})$ où L est la constante de Lipschitz définie en (7.7).

Pour $y_{i+1}^{(0)}$, on choisit une valeur y_{i+1} donnée par une méthode d'Adams-Bashforth qui joue le rôle de *méthode de prédiction*.

La méthode d'Adams-Moulton est une méthode de correction : elle corrige à chaque étape la solution donnée par la méthode explicite précédente. L'ensemble de deux méthodes, l'une de prédiction et l'autre de correction constitue une *méthode de prédiction-correction*.

7.3.3 Formulation générale des méthodes à pas multiples

Les méthodes à pas multiples précédentes entrent dans le cadre général des méthodes à k pas : calculer $\{y_i, i = 0, \dots, n\}$ solution des équations :

$$y_i = \eta_i(h) \quad \text{donné dans } \mathbb{R}, \quad i = 0, \dots, k-1 \tag{7.50}$$

$$\sum_{m=0}^{k} \alpha_m y_{m+i} = h \sum_{m=0}^{k} \beta_m f(x_{i+m}, y_{i+m}), \quad i = 0, \dots, n-k \tag{7.51}$$

L'entier k et les coefficients α_m et β_m pour $m = 0, \dots, k$ définissent la méthode. On suppose ici que $\alpha_k \neq 0$ et $|\alpha_0| + |\beta_0| > 0$:

— si $\beta_k = 0$, la méthode est *explicite* ;

— si $\beta_k \neq 0$, la méthode est *implicite*.

Ces méthodes ne sont pas auto-démarrantes ; les valeurs $y_1, y_2, \dots, y_{k-1}$ sont calculées par une méthode à un pas (de même ordre).

On définit comme précédemment des méthodes de prédiction-correction, une méthode implicite est couplée à une méthode explicite ; la méthode implicite est résolue par une méthode itérative [théorème 7.11] initialisée par la solution de la méthode explicite.

Théorème 7.12 Sous les hypothèses du théorème (7.2), si $\beta_k \neq 0$ et si :

$$h < \frac{|\alpha_k|}{\mathrm{L}|\beta_k|} \tag{7.52}$$

où L est la constante de Lipschitz de f, alors l'équation (7.51) admet une solution et une seule. Cette solution peut alors être calculée à l'aide de la méthode itérative :

$$\begin{cases} y_{i+k}^{(0)} & \text{calculé par la méthode de prédiction} \\ y_{i+k}^{(j+1)} = \dfrac{h\beta_k}{\alpha_k} f\left(x_{i+k}, y_{i+k}^{(j)}\right) + \left(\dfrac{h}{\alpha_k} \displaystyle\sum_{m=0}^{k-1} \beta_m f_{i+m} - \dfrac{1}{\alpha_k} \sum_{m=0}^{k-1} \alpha_m y_{i+m}\right) \end{cases} \tag{7.53}$$

Si (7.52) est vérifiée, la suite $(y_{i+k}^{(j)})$ converge vers y_{i+k} lorsque $j \to +\infty$.

Comme pour les méthodes à un pas, on peut introduire les notions de *consistance*, *convergence*, *stabilité* et *ordre* d'une méthode à pas multiples.

Ci-dessous, on indique des exemples de méthodes de prédiction-correction.

■ **Exemple — Méthode de Milne.** Il s'agit d'une méthode à trois pas et d'ordre 4 :

$$\begin{aligned} y_{i+1} &= y_{i-3} + \frac{h}{3}(8f_i - 4f_{i-1} + 8f_{i-2}) \\ y_{i+1} &= y_{i-1} + \frac{h}{3}(f_{i+1} + 4f_i + f_{i-1}) \end{aligned} \tag{7.54}$$

■ **Exemple — Méthode d'Adams.** Il s'agit d'une méthode à quatre pas et d'ordre 4 :

$$\begin{aligned} y_{i+1} &= y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}) \\ y_{i+1} &= y_i + \frac{h}{24}(9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}) \end{aligned} \tag{7.55}$$

8

Systèmes d'équations non linéaires

▷ **Problème 8.1** Étant donnée une application non linéaire $\mathbf{f}$, définie d'un ouvert U de $\mathbb{R}^n$ dans $\mathbb{R}^n$, on cherche le (ou les) $\mathbf{x} \in \mathrm{U}$ tel que :

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \tag{8.1}$$

8.1 Principe de résolution par itérations

On étudie des méthodes de résolution de (8.1) par itérations ; ceci consiste à déterminer une application :

$$\mathbf{g} : \mathrm{U} \subset \mathbb{R}^n \to \mathrm{U} \tag{8.2}$$

telle que :

1. $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ est équivalent à :
$$\mathbf{g}(\mathbf{x}) = \mathbf{x} \tag{8.3}$$
2. la suite $(\mathbf{x}^{(m)})$, $m \in \mathbb{N}$, définie par récurrence par :
$$\begin{cases} \mathbf{x}^{(0)} \text{ donné dans } \mathrm{U}' \subset \mathrm{U} \\ \mathbf{x}^{(m+1)} = \mathbf{g}(\mathbf{x}^{(m)}), \quad \text{pour } m \in \mathbb{N} \end{cases} \tag{8.4}$$
est convergente dans U.
3. $\lim \mathbf{x}^{(m)}$ quand $m \to \infty$ est une solution $\mathbf{x}$ de (8.1).

▷ **Définition 8.2 — Méthode itérative à un point.** On appelle *méthode itérative à un point* toute formule de la forme (8.4).

▷ **Définition 8.3 — Convergence.** Une méthode itérative à un point est dite *convergente* s'il existe un ouvert $\mathrm{U}' \subset \mathrm{U}$ tel que :

1. $\forall \mathbf{x}^{(0)} \in \mathrm{U}'$, $\lim \mathbf{x}^{(m)}$ existe quand $m \to \infty$;
2. $\mathbf{x} = \lim \mathbf{x}^{(m)}$ est solution de (8.1).

▷ **Définition 8.4 — Ordre.** Une méthode itérative convergente est dite *d'ordre* p s'il existe des constantes $c_1 > 0$ et $c_2 > 0$ telles que pour m assez grand :

$$c_1 \leqslant \frac{\|\mathbf{x}^{(m+1)} - \mathbf{x}\|}{\|\mathbf{x}^{(m)} - \mathbf{x}\|^p} \leqslant c_2 \tag{8.5}$$

avec $c_2 < 1$ lorsque $p = 1$.

Les démonstrations de convergence utilisent des théorèmes de point fixe. En particulier, nous avons :

Théorème 8.5 Lorsque $n = 1$, soit $g : \mathrm{I} \subset \mathbb{R} \to \mathbb{R}$, $g \in \mathcal{C}^1(\mathrm{I})$. Soit x une racine de $g(x) = x$, alors :

1. si $|g'(x)| < 1$, il existe un voisinage J de x dans I tel que :
   — l'équation (8.3) n'admet qu'une seule racine dans J ;
   — cette racine est la limite de le suite $x^{(m+1)} = g(x^{(m)})$ quel que soit le choix initial $x^{(0)} \in \mathrm{J}$.
2. si, de plus, $g'(x) \neq 0$, la méthode est d'ordre un ;
3. si $g \in \mathcal{C}^p(\mathrm{I})$ avec $p > 1$ et vérifie :

$$\begin{cases} g(x) = x, \\ g'(x) = \ldots = g^{(p-1)}(x) = 0, \\ g^{(p)}(x) \neq 0, \\ |g^{(p)}(x)| < p!\,\mathrm{K} \text{ avec } \mathrm{K} \in]0\,;1[, \end{cases} \tag{8.6}$$

alors la méthode est d'ordre p.

On a un résultat analogue dans $\mathbb{R}^n$ en remplaçant $|\cdot|$ par une norme de $\mathbb{R}^n$ et $g'(x)$ par la matrice gradient :

$$\mathbf{Dg(x)} = \left(\frac{\partial g_i(\mathbf{x})}{\partial x_j} \right)_{i,j=1,\ldots,n} \quad \text{où} \quad g = (g_i)_{i=1,\ldots,n} \tag{8.7}$$

et en remplaçant la condition $|g'(x)| < 1$ par :

$$\|\mathbf{Dg(x)}\| \leqslant \mathrm{K}/n \quad \text{avec} \quad \mathrm{K} < 1 \tag{8.8}$$

8.2 Principales méthodes en une dimension

Soit $f : \mathrm{I} \subset \mathbb{R} \to \mathbb{R}$.

8.2.1 Méthode des approximations successives

La fonction g s'écrit $g(y) = y - cf(y)$ où c est une constante non nulle choisie « convenablement ».

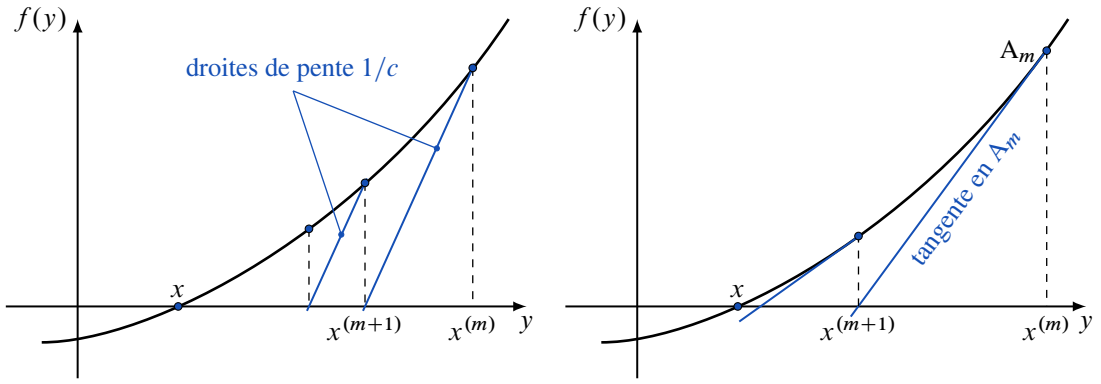
▷ **Proposition 8.6** Si $f \in \mathcal{C}^1(\mathrm{I}\,;\mathbb{R})$ et si $f'(x) \neq 0$ où x est une solution de $f(x) = 0$, alors la méthode itérative :

$$\begin{cases} x^{(0)} \text{ donné} \\ x^{(m+1)} = x^{(m)} - cf(x^{(m)}), \quad m \in \mathbb{N} \end{cases} \tag{8.9}$$

converge pour tout $c \in \mathbb{R}$ tel que

$$0 < cf'(x) < 2 \tag{8.10}$$

C'est une méthode du premier ordre. Cette méthode est illustrée sur la figure 8.1a.



(a) Méthode des approximations successives (b) Méthode de Newton-Raphson

Figure 8.1 – Interprétation géométrique

8.2.2 Méthode de Newton-Raphson

La fonction g est de la forme :

$$g(y) = y - \frac{f(y)}{f'(y)} \tag{8.11}$$

Le schéma itératif est alors le suivant :

$$\begin{cases} x^{(0)} \text{ donné} \\ x^{(m+1)} = x^{(m)} - \dfrac{f(x^{(m)})}{f'(x^{(m)})}, \quad \text{si } f'(x^{(m)}) \neq 0 \end{cases} \tag{8.12}$$

Il s'écrit de manière équivalente :

$$f(x^{(m)}) + (x^{(m+1)} - x^{(m)}) f'(x^{(m)}) = 0, \quad \text{si } f'(x^{(m)}) \neq 0 \tag{8.13}$$

Nous avons le résultat de convergence suivant :

▷ **Proposition 8.7** Si $f \in \mathcal{C}^2([a\,;b])$ vérifie :

1. $f(a)f(b) < 0$
2. $\forall y \in [a\,;b],\ f'(y) \neq 0$
3. $\forall y \in [a\,;b],\ f''(y) \neq 0$

alors, pour tout $x^{(0)} \in [a\,;b]$ tel que :

$$f(x^{(0)}) \cdot f''(x^{(0)}) > 0 \tag{8.14}$$

la suite $(x^{(m)})$ définie par (8.12) converge vers l'unique solution de $f(x) = 0$ dans $[a\,;b]$. La méthode est alors d'ordre 2. Cette méthode est illustrée sur la figure 8.1b.

Remarque Lorsque la racine x de l'équation $f(x) = 0$ est de multiplicité $p > 1$, la méthode itérative de Newton n'est plus que d'ordre un.

8.2.3 Méthode de dichotomie

Cette méthode prend aussi le nom de méthode de Bolzano. Soit $f : [a\,;b] \to \mathbb{R}$ continue. On suppose que $f(a)\,f(b) < 0$ et on pose :

$$\begin{cases} a_0 = a \\ b_0 = b \\ \text{pour } n \geqslant 1 : \quad c_n = (a_{n-1} + b_{n-1})/2 \\ \quad \text{si } f(a_{n-1})\,f(c_n) > 0, \text{ on pose } a_n = c_n \text{ et } b_n = b_{n-1} \\ \quad \text{si } f(a_{n-1})\,f(c_n) < 0, \text{ on pose } a_n = a_{n-1} \text{ et } b_n = c_n \end{cases} \tag{8.15}$$

▷ **Proposition 8.8** Sous les hypothèses $f : \mathcal{C}^0([a\,;b])$ et $f(a)\,f(b) < 0$, la suite (a_n) définie par (8.15) converge vers *une* racine de f dans $[a\,;b]$.

Remarque Sous les hypothèses précédentes, il peut y avoir plusieurs racines de f dans l'intervalle $[a\,;b]$ et la méthode de dichotomie assure la convergence vers l'une de ces racines.

8.3 Principales méthodes dans $\mathbb{R}^n$

8.3.1 Méthode des approximations successives

Le schéma est défini par une fonction $\mathbf{g}$ de la forme suivante :

$$\mathbf{g}(\mathbf{y}) = \mathbf{y} - \mathbf{A}\,\mathbf{f}(\mathbf{y}) \tag{8.16}$$

où $\mathbf{A} \in \mathbb{R}^{n,n}$ est une matrice inversible à coefficients constants.

▷ **Proposition 8.9** On suppose que $\mathbf{f}$ est continûment différentiable dans U. Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice telle que :

$$\|\mathbf{I} - \mathbf{A}\,\mathbf{Df}(\mathbf{x})\|_\infty < \frac{1}{n} \quad \text{où} \quad \mathbf{Df}(\mathbf{x}) = \left[\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right]_{i,j=1,\ldots,n} \tag{8.17}$$

et où $\mathbf{x}$ est une solution de $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, alors la méthode (8.16) converge dans un voisinage de $\mathbf{x}$.

8.3.2 Méthode de Newton

On remplace la matrice constante de (8.16) par :

$$\mathbf{A}(\mathbf{y}) \equiv \big[\mathbf{Df}(\mathbf{y})\big]^{-1} \quad \text{avec} \quad \mathbf{Df}(\mathbf{y}) = \left[\frac{\partial f_i(\mathbf{y})}{\partial y_j}\right]_{i,j=1,\ldots,n} \tag{8.18}$$

donc :

$$\mathbf{g}(\mathbf{y}) \equiv \mathbf{y} - \big[\mathbf{Df}(\mathbf{y})\big]^{-1}\mathbf{f}(\mathbf{y}) \tag{8.19}$$

Le schéma de Newton s'écrit :

$$\begin{cases} \mathbf{x}^{(0)} \text{ donné dans } \mathbb{R}^n \\ \mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \big[\mathbf{Df}(\mathbf{x}^{(m)})\big]^{-1}\mathbf{f}(\mathbf{x}^{(m)}) \end{cases} \tag{8.20}$$

dès que $\mathbf{Df}(\mathbf{x}^{(m)})$ est inversible. Le schéma s'écrit de manière équivalente :

$$\mathbf{f}(\mathbf{x}^{(m)}) + \mathbf{Df}(\mathbf{x}^{(m)})(\mathbf{x}^{(m-1)} - \mathbf{x}^{(m)}) = \mathbf{0} \tag{8.21}$$

La méthode est, en général, d'ordre 2.

8.4 Application aux racines de polynômes : méthode de Bairstow

Il s'agit d'une application au cas particulier où f est un polynôme. Étant donné un polynôme $\mathrm{P} \in \mathcal{P}_n$ de degré $n > 2$ à coefficients réels :

$$\mathrm{P}(x) = \sum_{i=0}^{n} a_i x^{n-i} = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n \tag{8.22}$$

on cherche toutes les racines de P, réelles ou complexes.

8.4.1 Principe de la méthode

La méthode de Bairstow permet de déterminer un trinôme $\mathrm{T}(x) = x^2 - sx + p$ (s et $p \in \mathbb{R}$) divisant P ainsi que les coefficients du polynôme quotient Q. Les racines du trinôme T réelles ou complexes, sont racines de P ; le procédé est alors itéré sur le polynôme quotient Q jusqu'à ce que ce polynôme soit de degré inférieur ou égal à 2.

Remarque On notera que si P est à coefficient réels, les calculs peuvent se faire en « réel », le calcul éventuel des racines complexes ne se faisant qu'en dernier lieu, lors du calcul des racines du trinôme T.

Pour s et p donnés quelconques, il existe $\mathrm{Q} \in \mathcal{P}_{n-2}$, $\mathrm{R} \in \mathbb{R}$ et $\mathrm{S} \in \mathbb{R}$ tels que :

$$\mathrm{P}(x) = (x^2 - sx + p)\mathrm{Q}(x) + \mathrm{R}x + \mathrm{S} \tag{8.23}$$

avec $\mathrm{Q}(x) = b_0 x^{n-2} + b_1 x^{n-3} + \cdots + b_{n-2}$. On cherche s et p tels que :

$$\begin{aligned} \mathrm{R} &= \mathrm{R}(s, p) = 0 \\ \mathrm{S} &= \mathrm{S}(s, p) = 0 \end{aligned} \tag{8.24}$$

Par identification, (et définition de b_{n-1} et b_n) nous avons :

$$\begin{cases} b_0 = a_0 \\ b_1 = a_1 + sb_0 \\ \quad \cdots\cdots\cdots \\ b_k = a_k + sb_{k-1} - pb_{k-2}, \quad k = 2, \ldots, n \end{cases} \tag{8.25}$$

avec :

$$\begin{cases} \mathrm{R} = b_{n-1} \\ \mathrm{S} = b_n - sb_{n-1} \end{cases} \tag{8.26}$$

Le système non linéaire (8.24) est résolu par la méthode itérative de Newton, sachant que les coefficients b_n et b_{n-1} du système dépendent des inconnues s et p par l'intermédiaire des équations (8.25).

Les dérivées partielles nécessaires au calcul sont obtenues par dérivation de (8.25) et (8.26) ; en particulier, si on note :

$$\alpha_k = \frac{\partial b_k}{\partial s}, \quad k = 0, \ldots, n \tag{8.27}$$

les coefficients α_k vérifient :

$$\begin{cases} \alpha_0 = 0 \\ \alpha_1 = b_0 \\ \quad \cdots\cdots\cdots \\ \alpha_k = b_{k-1} + s\alpha_{k-1} - p\alpha_{k-2}, \quad k = 2, \ldots, n \end{cases} \tag{8.28}$$

On vérifie d'autre part que les dérivées partielles :

$$\frac{\partial b_k}{\partial p},\ \frac{\partial \mathrm{R}}{\partial s},\ \frac{\partial \mathrm{R}}{\partial p},\ \frac{\partial \mathrm{S}}{\partial s},\ \frac{\partial \mathrm{S}}{\partial p} \tag{8.29}$$

s'expriment simplement par rapport aux coefficients α_k. II en résulte le schéma itératif de l'algorithme de Bairstow.

8.4.2 Algorithme

Étant donné un couple $(s^{(0)}, p^{(0)})$ de valeurs initiales. On définit par récurrence des couples $(s^{(m+1)}, p^{(m+1)})$ par :

- calcul des coefficients $b_i = b_i(s^{(m)}, p^{(m)})$ par (8.25) avec $s = s^{(m)}$ et $p = p^{(m)}$;
- calcul des coefficients $\alpha_i = \alpha_i(s^{(m)}, p^{(m)})$ par (8.28) avec $s = s^{(m)}$ et $p = p^{(m)}$;
- résolution du système linéaire donné par l'algorithme de Newton, c'est-à-dire :

$$\begin{cases} \alpha_{n-1}(s^{(m+1)} - s^{(m)}) - \alpha_{n-2}(p^{(m+1)} - p^{(m)}) + b_{n-1} = 0 \\ (\alpha_n - b_{n-1})(s^{(m+1)} - s^{(m)}) - \alpha_{n-1}(p^{(m+1)} - p^{(m)}) + b_n = 0 \end{cases} \tag{8.30}$$

La résolution du système (8.30) est possible dès que :

$$\Delta^{(m)} \equiv -\alpha_{n-1}^2 + \alpha_{n-2}(\alpha_n - b_{n-1}) \neq 0 \tag{8.31}$$

Par conséquent, la méthode de Bairstow est bien définie dès que :

$$\forall m \in \mathbb{N}, \quad \Delta^{(m)} \neq 0 \tag{8.32}$$

Les calculs des couples $(s^{(m)}, p^{(m)})$ sont itérés jusqu'à la convergence (éventuelle) des deux suites $(s^{(m)})$ et $(p^{(m)})$. Si s et p désignent les limites respectives des suites $s^{(m)}$ et $p^{(m)}$, deux racines de P sont obtenues par la résolution de l'équation du second degré :

$$x^2 - sx + p = 0 \tag{8.33}$$

9

Exercices

Exercice 1 — Formule du point milieu. On considère la formule du *point milieu* :

$$\int_a^b f(x)\,\mathrm{d}x \approx (b-a)\,f(m) \quad \text{où} \quad m = \frac{a+b}{2} \tag{9.1}$$

Montrer que cette formule est d'ordre 1.

Exercice 2 — Calcul de l'erreur de la formule des trapèzes.

1. Vérifier l'ordre de la méthode ;
2. Erreur de la formule des trapèzes : on suppose que $f \in \mathcal{C}^2([a\,;b])$. Définir une fonction $\phi \in \mathcal{C}^2([0\,;b-a])$ par :

$$\phi(t) = \int_a^{a+t} f(x)\,\mathrm{d}x - \frac{t}{2}(f(a)+f(a+t)) \quad \text{si } t \in [0\,;b-a] \tag{9.2}$$

Il faut montrer qu'il existe $\xi \in]a\,;b[$ tel que :

$$\phi(h) = -\frac{h^3}{12} f''(\xi) \tag{9.3}$$

— Montrer que $\phi(0) = \phi'(0) = \phi''(0) = 0$.
— On pose $\mathrm{Q}(t) = \phi(t) - \phi(h)t^3/h^3$. Montrer, à l'aide du théorème de Rolle, qu'il existe $t_0 \in]0\,;h[$ tel que $\mathrm{Q}''(t_0) = 0$. Conclure.

Exercice 3 — Calcul de l'erreur de la formule de Simpson.

1. Vérifier l'ordre de la méthode ;
2. Erreur de la formule de Simpson : on suppose que $f \in \mathcal{C}^4([a\,;b])$. Si $x_1 = (a+b)/2$, on pose, pour $t \in [\frac{a-b}{2}\,;\frac{b-a}{2}]$:

$$\phi(t) = \int_{x_1-t}^{x_1+t} f(x)\,\mathrm{d}x - \frac{t}{3}(f(x_1-t) + 4f(x_1) + f(x_1+t)) \tag{9.4}$$

Il faut montrer qu'il existe $\xi \in]a\,;b[$ tel que :

$$\phi(h) = -\frac{h^5}{90} f^{(4)}(\xi) \tag{9.5}$$

— Montrer que $\phi^{(i)}(0) = 0, i = 0, \ldots, 4$.

— On pose $Q(t) = \phi(t) - \phi(h)t^5/h^5$. Montrer, à l'aide du théorème de Rolle, qu'il existe $t_0 \in]0\,;h[$ tel que $Q^{(4)}(t_0) = 0$. Conclure.

Exercice 4 — Calcul de l'erreur de la formule du point milieu. On suppose que $f \in \mathcal{C}^2([a\,;b])$ et soit $x_0 = (b-a)/2$. On pose :

$$\phi(t) = \int_{x_0-t}^{x_0+t} f(x)\,\mathrm{d}x - 2tf(x_0) \tag{9.6}$$

Développer $\phi(t)$ à l'ordre 3 en $t = 0$ à l'aide du développement de Taylor-MacLaurin et conclure en considérant $\phi(h)$.

Exercice 5 — Formules composées du point milieu et des trapèzes. Considérons la formule du point milieu sur chacun des intervalles $[a_i\,;a_{i+1}]$, soit :

$$\mathrm{I}_n = h\sum_{i=0}^{n-1} f\Big(\frac{a_i + a_{i+1}}{2}\Big) \tag{9.7}$$

où $h = (b-a)/n$. On note $\mathrm{R}_n(f)$ l'erreur commise :

$$\mathrm{R}_n(f) = \int_a^b f(x)\,\mathrm{d}x - \mathrm{I}_n(f) \tag{9.8}$$

1. Montrer que si $f \in \mathcal{C}^2([a\,;b])$, alors il existe $\xi \in]a\,;b[$ tel que :
$$\mathrm{R}_n(f) = \frac{b-a}{24}h^2 f''(\xi) \tag{9.9}$$
2. En déduire que la méthode composée du point milieu est une méthode convergente.
3. Montrer que $\mathrm{I}_n(f)$ est une méthode d'intégration stable, c'est-à-dire que pour une perturbation δf bornée sur $[a\,;b]$ par ϵ donné, la différence en valeur absolue $|\mathrm{I}_n(f+\delta f) - \mathrm{I}_n(f)|$ reste bornée quand n augmente.
4. Mêmes questions pour la méthode composée de Simpson.

Exercice 6 — Résolution d'un problème aux limites en dimension 1 par la méthode des éléments finis. Étant donnée $f \in \mathcal{L}^2(]0\,;1[)$, on cherche une solution *régulière* u du problème suivant :

$$\begin{cases} -u'' = f \quad \text{dans} \quad [0\,;1] \\ u(0) = u(1) = 0 \end{cases} \tag{9.10}$$

On admettra, en dimension 1, que les fonctions $v \in \mathcal{H}^1(]0\,;1[)$ sont continues sur $[0\,;1]$ et donc que :

$$\mathcal{H}_0^1(]0\,;1[) = \{v \in \mathcal{H}^1(]0\,;1[)\,;\, v(0) = v(1) = 0\} \tag{9.11}$$

Le problème (9.10) admet une et une seule solution u dans $\mathcal{H}_0^1(]0\,;1[)$. Dans la suite, nous noterons $\mathrm{V} = \mathcal{H}_0^1(]0\,;1[)$ cet espace. On remarque que $f \in \mathcal{L}^2(]0\,;1[)$ entraîne $u'' \in \mathcal{L}^2(]0\,;1[)$ et donc $u \in \mathcal{H}^2(]0\,;1[)$. Il en résulte qu'alors $u \in \mathcal{C}^1([0\,;1])$.

1. Formulation variationnelle : montrer que la solution u de (9.10) dans $\mathcal{H}_0^1(]0\,;1[)$ est solution du problème variationnel :
$$u \in \mathrm{V} \quad \text{et} \quad \forall v \in \mathrm{V}, \quad \int_0^1 u'v'\,\mathrm{d}x = \int_0^1 fv\,\mathrm{d}x \tag{9.12}$$
Montrer que la réciproque est vraie pour u tel que $u'' \in \mathcal{L}^2(]0\,;1[)$.

2. Approximation de (9.12) par la méthode des éléments finis P_1 : on remplace V par un sous-espace V_h de dimension finie $n-1$ de polynômes par morceaux de degré inférieur ou égal à 1 associé à la subdivision uniforme $\{x_i = ih; i = 0, \ldots, n\}$, de pas $h = 1/n$:

$$V_h = \{v_h \in \mathcal{C}^0([0\,;1]); v_{h|[x_i;x_{i+1}]} \in P_1 \text{ et } v_h(0) = v_h(1) = 0\} \tag{9.13}$$

où P_1 désigne l'espace des polynômes de degré inférieur ou égal à 1. On recherche alors u_h solution du problème approché :

$$u_h \in V_h \quad \text{et} \quad \forall v_h \in V_h, \quad \int_0^1 u_h' v_h' \,\mathrm{d}x = \int_0^1 f v_h \,\mathrm{d}x \tag{9.14}$$

Pour cela, on considère la base $\{\phi_i\,; i = 1, \ldots, n-1\}$ de V_h dont un des éléments est illustré sur la figure 9.1. Montrer que (9.12) s'écrit sous la forme d'un système linéaire d'ordre $n-1$

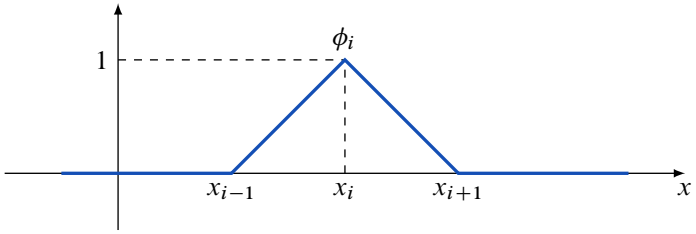


Figure 9.1 – Vecteur de base de V_h

dont la matrice est tridiagonale, symétrique et définie positive.

Erreur d'approximation On peut montrer l'existence de constantes C_1, C_2 et h_0 telles que :

$$\forall h \in]0\,;h_0], \quad \forall x \in [0\,;1], \quad |u(x) - u_h(x)| \leqslant C_1 h^{3/2} ||f||_{\mathcal{L}^2(]0;1[)} \tag{9.15}$$

et :

$$\forall h \in]0\,;h_0], \quad ||u - u_h|| \leqslant C_2 h^2 ||f||_{\mathcal{L}^2(]0;1[)} \tag{9.16}$$

où u est la solution dans V de (9.12) et u_h est la solution de (9.14).

Voir aussi l'annexe A où l'on donne des résultats d'erreur d'approximation lorsque l'on utilise une méthode d'intégration numérique pour calculer le produit scalaire :

$$(f, v_h)_0 = \int_0^1 f(x) v_h(x) \,\mathrm{d}x \tag{9.17}$$

Exercice 7 — Résolution d'un problème aux limites en dimension 1 par une méthode de différences finies.

1. Montrer que si f est une fonction $v \in \mathcal{C}^4([0\,;1])$ pour tout $x \in]0\,;1[$ et tout $h > 0$ tels que $]x-h\,;x+h[\subset [0\,;1]$, il existe $\xi \in]x\,;x+h[$, ξ_1 et $\xi_2 \in]x-h\,;x+h[$ tels que :

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2} f^{(2)}(\xi) \tag{9.18}$$

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6} f^{(3)}(\xi_1) \tag{9.19}$$

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{h^2}{12} f^{(4)}(\xi_2) \tag{9.20}$$

2. Étant donnée une fonction g continue sur $[0\,;1]$, on se propose de calculer une valeur approchée de la solution $y \in \mathcal{C}^2([0\,;1])$ du problème suivant :

$$\begin{cases} -y'' + y = g \quad \text{dans }]0\,;1[\\ y(0) = y(1) = 0 \end{cases} \tag{9.21}$$

Pour un $n \in \mathbb{N}$ donné, on définit une subdivision régulière $(x_i)_{i=0,\dots,n}$ de l'intervalle $[0\,;1]$ par $x_i = ih$, $i = 0,\dots,n$ avec $nh = 1$. Pour $i = 0,\dots,n$, on notera y_i une valeur approchée de $y(x_i)$.

(a) Écrire un système linéaire tridiagonal vérifié par les y_i pour $i = 1,\dots,n-1$ en approchant $y''(x_i)$ par le rapport

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} \tag{9.22}$$

(b) Montrer que la matrice du système linéaire obtenu est définie positive et que le système est inversible. Quelles méthodes numériques peut-on utiliser pour la résolution numérique de ce système ?

Remarque On peut montrer que si la solution y appartient à $\mathcal{C}^4([0\,;1])$ alors il existe deux constantes $C > 0$ et $h_0 > 0$ telles que pour tout $h \in]0\,;h_0]$:

$$\sup_{i=1,\dots,n} |y(x_i) - y_i| \leqslant Ch^2 \sup_{x\in[0;1]} |y^{(4)}(x)| \tag{9.23}$$

(c) Écrire un schéma de résolution numérique pour le problème suivant :

$$\begin{cases} -y'' + y = g \quad \text{dans }]0\,;1[\\ y(0) = \alpha \\ y(1) = \beta \end{cases} \tag{9.24}$$

avec α et β, réels.

Exercice 8 — Résolution d'un problème aux limites en dimension 2 par une méthode de différences finies. Soit $D =]0\,;1[\times]0\,;\ell[$, un rectangle de $\mathbb{R}^2$ avec $\ell > 0$, de frontière ∂D et soit f, une fonction donnée continue sur D. On considère le problème suivant : trouver u telle que :

$$\begin{cases} -\Delta u + u = f \quad \text{dans} \quad D \\ u(x,y) = 0 \qquad \text{sur} \quad \partial D \end{cases} \tag{9.25}$$

Ce problème admet une et une seule solution $u \in \mathcal{H}_0^1(D)$. On admettra que si $f \in \mathcal{C}^0(\bar{D})$ alors $u \in \mathcal{C}^0(\bar{D})$.

1. Trouver des formules analogues à (9.22) pour les dérivées partielles :

$$\frac{\partial^2 u}{\partial x^2}(x,y) \quad \text{et} \quad \frac{\partial^2 u}{\partial y^2}(x,y) \tag{9.26}$$

lorsque u est une fonction $u \in \mathcal{C}^4(\bar{D})$.

2. Schéma de résolution numérique de (9.25) : étant donnés deux entiers N et $P \in \mathbb{N}$, on considère la discrétisation suivante du domaine D. On définit un maillage régulier :

$$\{(x_i, y_j);\ x_i = ih, y_j = jg;\ i = 0,\dots,N,\ j = 0,\dots,P\} \tag{9.27}$$

où $h = 1/N$ et $g = 1/P$ sont les *pas de discrétisation* et on cherche une approximation de la solution u aux points de ce maillage. Pour $i = 0,\dots,N$, $j = 0,\dots,P$, on notera u_{ij} approximation de $u(x_i, y_j)$.

— Utiliser la question précédente pour définir une approximation de $\Delta u(x_i, y_j)$ pour $i = 0, \ldots, \mathrm{N}$, $j = 0, \ldots, \mathrm{P}$ par une relation liant les valeurs de u au point (x_i, y_j) et à quatre points voisins (schéma à cinq points).

— En déduire un système linéaire approché vérifié par des u_{ij} pour les indices $i = 1, \ldots, \mathrm{N}-1$ et $j = 1, \ldots, \mathrm{P}-1$. Les approximations de u aux points du bord du domaine doivent tenir compte des conditions aux limites. Écrire ce système par blocs. Pour cela, on pourra considérer les vecteurs blocs :

$$\mathrm{U} = \begin{pmatrix} \mathrm{U}_1 \\ \vdots \\ \mathrm{U}_j \\ \vdots \\ \mathrm{U}_{\mathrm{P}-1} \end{pmatrix} \quad \text{où} \quad \mathrm{U}_j = \begin{pmatrix} u_{1,j} \\ \vdots \\ u_{i,j} \\ \vdots \\ u_{\mathrm{N}-1,j} \end{pmatrix} \quad j = 1, \ldots, \mathrm{P}-1 \tag{9.28}$$

et les décompositions correspondantes de la matrice du système et du second membre.

— Vérifier que ce système est un système symétrique, bande, tridiagonal par blocs, défini positif. En déduire l'existence d'une approximation (u_{ij}) de u aux points (x_i, y_j) pour $i = 0, \ldots, \mathrm{N}$ et $j = 0, \ldots, \mathrm{P}$ par le schéma précédent.

Remarque On peut montrer que si $u \in \mathcal{C}^4(\bar{\mathrm{D}})$, il existe des constantes C, $h_0 > 0$ et $g_0 > 0$ telles que si $h \in]0\,; h_0]$ et $g \in]0\,; g_0]$:

$$\sup_{\substack{i=0,\ldots,\mathrm{N} \\ j=0,\ldots,\mathrm{P}}} |u(x_i, y_j) - u_{ij}| \leqslant \mathrm{C}(h^2 + g^2) \tag{9.29}$$

Exercice 9 — Normes matricielles. Pour $1 \leqslant p \leqslant +\infty$, on notera $||\mathbf{x}||_p$ la norme vectorielle définie sur $\mathbb{C}^n$ par :

$$\begin{aligned} ||\mathbf{x}||_p &= \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p} \quad \text{si} \quad 1 \leqslant p \leqslant +\infty \\ ||\mathbf{x}||_p &= \max_{i=1,\ldots,n} |x_i|^p \quad \text{si} \quad p = +\infty \end{aligned} \tag{9.30}$$

où $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x} = (x_i)_{i=1,\ldots,n}$. On propose de montrer que les normes matricielles sur $\mathbb{C}^{n,n}$ subordonnées aux normes vectorielles $||\mathbf{x}||_1$, $||\mathbf{x}||_2$ et $||\mathbf{x}||_\infty$ vérifient [proposition (1.15)] :

$$||\mathbf{A}||_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \tag{9.31}$$

$$||\mathbf{A}||_2 = \rho(\mathbf{A}^* \mathbf{A})^{1/2} \tag{9.32}$$

$$||\mathbf{A}||_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \tag{9.33}$$

où $\mathbf{A} \in \mathbb{C}^{n,n}$, $\mathbf{A} = (a_{ij})_{i,j=1,\ldots,n}$.

1. Pour montrer (9.31), soit :

$$\mathrm{M} = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \tag{9.34}$$

(a) Par majoration directe, montrer que :

$$\forall \mathbf{x} \in \mathbb{C}^n \text{ tel que } ||\mathbf{x}||_1 = 1 : \quad ||\mathbf{A}\mathbf{x}||_1 \leqslant \mathrm{M} \tag{9.35}$$

(b) En déduire que :

$$||\mathbf{A}||_1 \leqslant \mathrm{M} \tag{9.36}$$

(c) Construire $\mathbf{y} \in \mathbb{R}^n$ tel que :

$$||\mathbf{y}||_1 = 1 \ \text{ et } \ ||\mathbf{A}\mathbf{y}||_1 = \mathrm{M} \tag{9.37}$$

(d) Conclure.

2. Pour (9.32), montrer que :
   (a) Pour toute matrice $\mathbf{A} \in \mathbb{C}^{n,n}$, la matrice $\mathbf{A}^* \mathbf{A}$ est hermitienne et positive (si $\mathbf{A} \in \mathbb{R}^{n,n}$, la matrice $\mathbf{A}^* \mathbf{A}$ est alors réelle, symétrique et positive). On rappelle qu'une matrice hermitienne (ou une matrice réelle, symétrique) admet une base orthonormée de vecteurs propres.
   (b) Montrer que les valeurs propres de $\mathbf{A}^* \mathbf{A}$ sont réelles et positives. Utiliser alors la méthode de (9.31) en exprimant un vecteur $\mathbf{x}$ sur une base orthonormée de vecteurs propres.
3. Pour montrer (9.33), utiliser la méthode de (9.31). Lorsque $\mathbf{A}$ est réelle, considérer le vecteur :

$$\mathbf{y} = (y_i), \quad \begin{cases} y_i = 1 & \text{si } a_{ki} \geqslant 0 \\ y_i = -1 & \text{si } a_{ki} < 0 \end{cases} \tag{9.38}$$

où k est un indice tel que :

$$\sum_{j=1}^{n} |a_{kj}| = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \tag{9.39}$$

Adapter la preuve pour le cas complexe.

4. Montrer que si $\mathbf{A}$ est une matrice hermitienne (ou bien réelle symétrique) alors :

$$||\mathbf{A}||_2 = \rho(\mathbf{A}) \tag{9.40}$$

Exercice 10 — Conditionnement d'une matrice. Soit $||\cdot||$ une norme matricielle sur $\mathbb{R}^{n,n}$ et $||\cdot||$, une norme vectorielle compatible sur $\mathbb{R}^n$, c'est-à-dire telle que :

$$\forall \mathbf{A} \in \mathbb{R}^{n,n}, \quad \forall \mathbf{x} \in \mathbb{R}^n : \quad ||\mathbf{A}\mathbf{x}|| \leqslant ||\mathbf{A}||\,||\mathbf{x}|| \tag{9.41}$$

Si $\mathbf{A} \in \mathbb{R}^{n,n}$ est une matrice *inversible*, on définit le *conditionnement* de $\mathbf{A}$ relativement à la norme $||\cdot||$, et on note $\mathrm{cond}(\mathbf{A})$, le nombre :

$$\mathrm{cond}(\mathbf{A}) = ||\mathbf{A}||\,||\mathbf{A}^{-1}|| \tag{9.42}$$

1. Montrer que lorsque la norme matricielle choisie vérifie :

$$||\mathbf{I}|| = 1 \tag{9.43}$$

où $\mathbf{I}$ est la matrice identité, alors :

$$\mathrm{cond}(\mathbf{A}) \geqslant 1 \tag{9.44}$$

Donner des exemples de normes vérifiant (9.43). On dira alors que la matrice $\mathbf{A}$ est *bien conditionnée* si son conditionnement est voisin de 1 et *mal conditionnée* si son conditionnement est « grand ». Plus $\mathrm{cond}(\mathbf{A})$ sera petit, meilleur sera le conditionnement de $\mathbf{A}$.

2. On considère la norme matricielle $||\cdot||_2$ et on note $\text{cond}_2(\cdot)$ le conditionnement correspondant. Montrer que pour pour toute matrice carrée $\mathbf{Q} \in \mathbb{R}^{n,n}$ orthogonale ($\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$), on a :

$$\text{cond}_2(\mathbf{Q}) = 1 \tag{9.45}$$

3. Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice *inversible* et $\mathbf{b} \in \mathbb{R}^n$ un vecteur non nul. Soit $\mathbf{x}$ la solution du système linéaire $\mathbf{Ax} = \mathbf{b}$ et $\mathbf{x} + \boldsymbol{\delta}\mathbf{x}$ celle du système $\mathbf{Ay} = \mathbf{b} + \boldsymbol{\delta}\mathbf{b}$. Montrer que :

$$\frac{||\boldsymbol{\delta}\mathbf{x}||}{||\mathbf{x}||} \leqslant \text{cond}(\mathbf{A}) \frac{||\boldsymbol{\delta}\mathbf{b}||}{||\mathbf{b}||} \tag{9.46}$$

Application. Soit :

$$\mathbf{A} = \begin{bmatrix} 0{,}5 & 0{,}4 \\ 0{,}3 & 0{,}25 \end{bmatrix} \quad \text{et} \quad \mathbf{b} = \begin{pmatrix} 0{,}2 \\ 1 \end{pmatrix} \tag{9.47}$$

Donner une estimation de $||\boldsymbol{\delta}\mathbf{x}||/||\mathbf{x}||$ pour la norme $||\cdot||_\infty$ lorsque :

$$\boldsymbol{\delta}\mathbf{b} = \begin{pmatrix} 0 \\ 10^{-4} \end{pmatrix} \tag{9.48}$$

4. On suppose maintenant que $\mathbf{A}$ et $\mathbf{A} + \boldsymbol{\delta}\mathbf{A}$ sont inversibles. Soit $\mathbf{x}$ la solution du système linéaire $\mathbf{Ax} = \mathbf{b}$ et $\mathbf{x} + \boldsymbol{\delta}\mathbf{x}$ celle du système $(\mathbf{A} + \boldsymbol{\delta}\mathbf{A})\mathbf{y} = \mathbf{b}$. Montrer que :

$$\frac{||\boldsymbol{\delta}\mathbf{x}||}{||\mathbf{x} + \boldsymbol{\delta}\mathbf{x}||} \leqslant \text{cond}(\mathbf{A}) \frac{||\boldsymbol{\delta}\mathbf{A}||}{||\mathbf{A}||} \tag{9.49}$$

Application. Soit :

$$\mathbf{A} = \begin{bmatrix} 1 & 10^3 \\ 10^3 & 10^6 + 10^3 \end{bmatrix} \quad \text{et} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{9.50}$$

Donner une estimation de $||\boldsymbol{\delta}\mathbf{x}||/||\mathbf{x} + \boldsymbol{\delta}\mathbf{x}||$ pour la norme $||\cdot||_\infty$ lorsque :

$$\boldsymbol{\delta}\mathbf{A} = \begin{pmatrix} 10^{-3} & 0 \\ 0 & 0 \end{pmatrix} \tag{9.51}$$

5. Matrice $\mathbf{A}$ réelle, symétrique et inversible. Soient $|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda_n| > 0$, les valeurs propres (réelles) de $\mathbf{A}$ rangées dans l'ordre décroissant des valeurs absolues. Montrer que, pour la norme $||\cdot||_2$, on a :

$$\text{cond}_2(\mathbf{A}) = \frac{|\lambda_1|}{|\lambda_n|} \tag{9.52}$$

Exercice 11 — Factorisation LU ou LDR d'une matrice A.

1. Montrer qu'une matrice $\mathbf{A}$ d'ordre n admet une factorisation $\mathbf{A} = \mathbf{LU}$ où $\mathbf{L}$ est triangulaire inférieure à diagonale unité et $\mathbf{U}$ triangulaire supérieure, si et seulement si toutes les sous-matrices principales de $\mathbf{A}$ sont régulières ;
2. Montrer qu'alors les matrices $\mathbf{L}$ et $\mathbf{U}$ sont uniques ;
3. Montrer que la factorisation $\mathbf{A} = \mathbf{LU}$ peut aussi s'écrire $\mathbf{A} = \mathbf{LDR}$ où $\mathbf{L}$ est triangulaire inférieure à diagonale unité, $\mathbf{D}$ diagonale et $\mathbf{R}$ triangulaire supérieure à diagonale unité.
4. Montrer que si de plus $\mathbf{A}$ est réelle, symétrique, alors la factorisation $\mathbf{A} = \mathbf{LU}$ peut s'écrire $\mathbf{A} = \mathbf{LDL}^\top$.

Exercice 12 — Factorisation de Cholesky : $\mathbf{A} = \mathbf{S}\mathbf{S}^\top$.

1. Montrer que si $\mathbf{A}$ est réelle, symétrique, définie positive, alors ses sous-matrices principales sont aussi symétriques, définies positives et leurs déterminants sont strictement positifs.
2. En déduire qu'une matrice $\mathbf{A}$ réelle, symétrique, définie positive admet une factorisation $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^\top$ qui peut s'écrire $\mathbf{A} = \mathbf{S}\mathbf{S}^\top$ où $\mathbf{S}$ est triangulaire inférieure, cette factorisation étant unique au signe près des coefficients de la diagonale de $\mathbf{S}$ (tous positifs ou tous négatifs).

Exercice 13 — Méthodes itératives de résolution numérique d'un système linéaire. Montrer la convergence des méthodes itératives de Jacobi et de Gauss-Seidel pour la résolution d'un système linéaire :

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{où} \quad \mathbf{A} \in \mathbb{R}^{n,n} \ \text{ et } \ \mathbf{b} \in \mathbb{R}^n \tag{9.53}$$

lorsque la matrice $\mathbf{A}$ est à diagonale strictement dominante. Pour chacune de ces méthodes, on pourra considérer le vecteur erreur d'approximation $\mathbf{e}^{(p)} = \mathbf{x}^{(p)} - \mathbf{x}$ à la p^{e} étape de composantes $\mathbf{e}^{(p)} = ({e_i}^{(p)})_{i=1,\dots,n}$:

1. montrer qu'il existe un réel $r \in [0\,;1[$ tel que :

$$\forall i = 1,\dots,n, \quad |{e_i}^{(p)}| \leqslant r||\mathbf{e}^{(p-1)}||_\infty \tag{9.54}$$

2. établir la limite :

$$\lim_{p\to+\infty} ||\mathbf{e}^{(p)}||_\infty = 0 \tag{9.55}$$

Exercice 14 — Factorisations de matrices réelles, symétriques et définies positives.

1. Le double but est d'étudier diverses décompositions d'une matrice réelle, symétrique, définie positive et de construire une méthode numérique de résolution d'un système linéaire adapté au cas d'une telle matrice.
   (a) Soit $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n,n}$ une matrice réelle, symétrique, définie positive. Montrer que les sous-matrices principales $\mathbf{A}_k \in \mathbb{R}^{k,k}$ de $\mathbf{A}$ ($k \in \{1,\dots,n\}$) de coefficients $(\mathbf{A}_k)_{ij} = a_{ij}$ (pour $i, j = 1,\dots,k$), sont aussi symétriques, définies positives. En déduire que $\det(\mathbf{A}_k) > 0$ pour $k = 1,\dots,n$.
   (b) Soit $\mathbf{A} \in \mathbb{R}^{n,n}$. Écrire des relations reliant les coefficients de $\mathbf{A}$ et ceux de $\mathbf{S}$, qui soient équivalentes aux conditions d'existence d'une matrice triangulaire inférieure $\mathbf{S}$ telle que $\mathbf{A} = \mathbf{S}\mathbf{S}^\top$.
   (c) Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice réelle, symétrique, définie positive. Construire une matrice $\mathbf{S}$ triangulaire inférieure inversible telle que $\mathbf{A} = \mathbf{S}\mathbf{S}^\top$. On pourra le faire colonne par colonne (par récurrence sur l'indice de colonne) ou encore ligne par ligne, en utilisant les résultats de 1. Vérifier que l'on peut *choisir* les coefficients diagonaux de $\mathbf{S}$ strictement positifs ; montrer que, avec ce choix, on a unicité de la décomposition.
   (d) Montrer la réciproque : si $\mathbf{A}$ est une matrice réelle admettant une factorisation $\mathbf{A} = \mathbf{S}\mathbf{S}^\top$ où $\mathbf{S}$ est triangulaire inférieure inversible, alors $\mathbf{A}$ est symétrique, définie positive.
   (e) Déduire de (c) et (d) une méthode de résolution d'un système linéaire $\mathbf{A}\mathbf{x} = \mathbf{b}$ lorsque $\mathbf{A}$ est une matrice réelle, symétrique, définie positive. Quel est le nombre d'opérations nécessitées par la méthode appelée *méthode de Cholesky*. Comparer avec la méthode de Gauss et la méthode de Cramer.
   (f) Donner un *test numérique* permettant d'affirmer qu'une matrice réelle symétrique est, ou n'est pas, définie positive.
   (g) Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice réelle, symétrique, définie positive. On suppose ici de plus que $\mathbf{A}$ est $(2\ell+1)$ diagonale. Montrer qu'alors la matrice triangulaire $\mathbf{S}$ construite en (c) est une matrice $(\ell+1, 1)$.

2. On suppose toujours que $\mathbf{A}$ est une matrice réelle, symétrique, définie positive. Montrer que $\mathbf{A}$ admet une factorisation $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^\top$, où $\mathbf{L}$ est une matrice triangulaire inférieure à diagonale unité et $\mathbf{D}$ une matrice diagonale à coefficients strictement positifs.
3. En déduire qu'une matrice $\mathbf{A}$ réelle, symétrique, définie positive admet une factorisation $\mathbf{A} = \mathbf{L}\mathbf{U}$ sans permutations ni de lignes, ni de colonnes, où $\mathbf{L}$ est une matrice triangulaire inférieure à diagonale unité et $\mathbf{U}$, une matrice triangulaire supérieure inversible.

Exercice 15 — Méthode itérative de résolution numérique d'un système linéaire. Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice réelle, symétrique et définie positive. On note :

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n \tag{9.56}$$

ses valeurs propres classées dans l'ordre décroissant. Soit $\mathbf{b}$ un vecteur de $\mathbb{R}^n$ et α, un réel non nul. Pour résoudre le système linéaire :

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{9.57}$$

on considère la méthode itérative suivante :

$$\begin{cases} \mathbf{x}^{(0)} \quad \text{donné dans } \mathbb{R}^n \\ \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}), \quad k \in \mathbb{N}, \quad \mathbf{x}^{(k)} \in \mathbb{R}^n \end{cases} \tag{9.58}$$

1. À quelle décomposition $\mathbf{A} = \mathbf{M} - \mathbf{N}$ correspond la méthode itérative (9.58) ? Montrer que cette méthode itérative converge si et seulement si :

$$0 < \alpha < \frac{2}{\lambda_1} \tag{9.59}$$

2. Vérifier que le rayon spectral de $\mathbf{M}^{-1}\mathbf{N}$ est donné par :

$$\rho(\mathbf{M}^{-1}\mathbf{N}) = \max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\} \tag{9.60}$$

3. Tracer le graphe de la fonction :

$$\alpha \in \mathbb{R} \to \max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\} \tag{9.61}$$

et montrer que le choix optimal de α est donné par :

$$\alpha = \frac{2}{\lambda_1 + \lambda_n} \tag{9.62}$$

Exercice 16 — Vecteurs propres. Déterminer les vecteurs propres d'une matrice triangulaire inférieure lorsque les éléments diagonaux sont distincts deux à deux.

Exercice 17 — Calcul numérique de valeurs propres. Démontrer le théorème 3.6 donnant la convergence de la méthode du quotient de Rayleigh pour le calcul de la plus grande (en valeur absolue) valeur propre d'une matrice réelle symétrique. On décomposera le vecteur initial sur une base orthonormée de vecteurs propres.

Exercice 18 — Approximation de fonctions. Déterminer *le* polynôme p_2 de degré inférieur ou égal à 2 qui réalise la meilleure approximation de la fonction $f(x) = \cos(x)$ sur $[-\pi/2\,;\pi/2]$ pour la norme de $\mathcal{L}^2_w(]-\pi/2\,;\pi/2[)$ définie par le produit scalaire :

$$\forall u, v \in \mathcal{L}^2_w(]-\pi/2\,;\pi/2[), \quad \langle u, v\rangle_w = \int_{-\pi/2}^{\pi/2} w(x)u(x)v(x)\,\mathrm{d}x \tag{9.63}$$

avec $w(x) = 1$ sur $[-\pi/2\,;\pi/2]$. On utilisera deux méthodes :

1. en construisant une base orthogonale de $\mathcal{P}_2$, polynômes de degré inférieur ou égal à 2 ;
2. en utilisant la base de $\mathcal{P}_2$ constituée des monômes $\{x^i,\ i = 0, 1, 2\}$.

Exercice 19 — Polynômes orthogonaux. Soit w une fonction poids positive sur un intervalle $]a\,;b[$ de $\mathbb{R}$ telle que :

$$\forall n \in \mathbb{N}, \quad \int_a^b w(x)x^n \,\mathrm{d}x < +\infty \tag{9.64}$$

1. Montrer l'existence d'une suite de polynômes $p_0,\ p_1, \ldots, p_n, \ldots$ vérifiant :

$$\begin{cases} \deg(p_n) = n \\ \forall q \in \mathcal{P}_{n-1}, \quad \langle p_n, q\rangle_w = 0 \end{cases} \tag{9.65}$$

   où $\langle \cdot, \cdot \rangle_w$ désigne le produit scalaire défini sur $\mathcal{L}^2_w(]a\,;b[)$ par :

$$\langle u, v\rangle_w = \int_a^b w(x)u(x)v(x)\,\mathrm{d}x \tag{9.66}$$

   On désignera par $||\cdot||_w$, la norme associée. Montrer l'unicité, à une constante multiplicative près, de cette suite de polynômes.
2. On suppose maintenant que la suite (p_n) construite en 1. est unitaire. Montrer qu'alors elle vérifie la relation de récurrence :

$$p_n(x) = (x - \alpha_n)p_{n-1}(x) - \beta_n p_{n-2}(x) \tag{9.67}$$

   où :

$$\alpha_n = \frac{\langle xp_{n-1}, p_{n-1}\rangle_w}{||p_{n-1}||^2_w} \quad ; \quad \beta_n = \frac{||p_{n-1}||^2_w}{||p_{n-2}||^2_w} \tag{9.68}$$

   On pourra remarquer que $p_n - xp_{n-1}$ est un polynôme de degré inférieur ou égal à $n - 1$; calculer ses coefficients sur la base (p_j) en effectuant le produit scalaire de ce polynôme avec chacun des p_j pour $j = 0, \ldots, n - 1$ et en déduire la formule de récurrence (9.67)-(9.68).
3. Donner une base orthonormée de l'espace vectoriel des polynômes pour le produit scalaire $\langle \cdot, \cdot \rangle_w$.
4. Montrer que les n racines du polynôme p_n sont réelles, distinctes et intérieures à l'intervalle $]a\,;b[$.

Exercice 20 — Intégration numérique. Soit $\alpha \in [-1\,;1]$. Déterminer l'ordre des formules d'intégration numérique suivantes en fonction des valeurs du paramètre α :

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \big(f(-\alpha) + f(\alpha)\big) + \mathrm{R}_{[-1;1]}(f), \quad f \in \mathcal{C}^0([-1\,;1]) \tag{9.69}$$

Exercice 21 — Intégration numérique sur un carré.

1. En partant de la formule de Gauss à deux points :

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx f(-1/\sqrt{3}) + f(1/\sqrt{3}) \tag{9.70}$$

   retrouver la formule de Gauss-Legendre (6.38) et vérifier qu'elle est exacte sur Q_3.

2. De façon analogue, en partant de la formule de Gauss-Legendre en une dimension à trois points sur $[-1\,;1]$, donner une formule à deux dimensions sur le carré qui soit exacte sur Q_5.

Exercice 22 — Intégration numérique sur un triangle.

1. Montrer que la formule de Gauss à un point (6.41) est exacte sur P_1.
2. Retrouver cette formule en calculant λ_0, x_0 et y_0 tels que la formule $\mathrm{I}_0 = \lambda_0 f(x_0, y_0)$ soit exacte sur P_1.
3. Montrer que la formule de Gauss à trois points (6.42) est exacte sur P_3.

Exercice 23 — Polynômes de Tchebychev. On définit sur $\mathcal{L}^2_w(]-1\,;1[)$ le produit scalaire :

$$\forall u,\, v \in \mathcal{L}^2_w(]-1\,;1[), \quad \langle u, v\rangle_w = \int_{-1}^{1} u(x)v(x)\frac{\mathrm{d}x}{\sqrt{1-x^2}} \tag{9.71}$$

1. Montrer, en utilisant des résultats du cours ou de l'exercice 19, l'existence d'une famille *unique* $(\mathrm{T}_n)_{n\in\mathbb{N}}$ de polynômes orthogonaux pour le produit scalaire (9.71) et tels que $\deg(\mathrm{T}_n) = n$ et $\mathrm{T}_n(1) = 1$ pour $n \in \mathbb{N}$.
2. On considère la fonction :

$$\forall x \in [-1\,;1], \quad \mathrm{F}_n(x) = \cos(n \arccos x) \tag{9.72}$$

Montrer la relation de récurrence :

$$n \geqslant 1, \quad x \in [-1\,;1], \quad \mathrm{F}_{n+1}(x) = 2\mathrm{F}_1(x)\mathrm{F}_n(x) - \mathrm{F}_{n-1}(x) \tag{9.73}$$

En déduire que :

$$n \geqslant 0, \quad y \in [0\,;\pi], \quad \mathrm{T}_n(\cos(y)) = \cos(ny) \tag{9.74}$$

Quelles sont les racines de T_n ?

3. Calculer $\langle \mathrm{T}_n, \mathrm{T}_n\rangle$ pour tout $n \in \mathbb{N}$.
4. Calculer le coefficient du terme de plus haut degré de T_n.
5. Exprimer les fonctions 1, x, x^2 et x^3 comme combinaisons linéaires des polynômes T_0, T_1, T_2 et T_3.
6. Écrire une méthode d'intégration numérique de la forme :

$$\int_{-1}^{1} \frac{f(x)\,\mathrm{d}x}{\sqrt{1-x^2}} = \sum_{i=0}^{\mathrm{N}} \lambda_i f(x_i) \quad \text{pour } f \in \mathcal{C}^0([-1\,;1]) \tag{9.75}$$

qui soit exacte pour tout polynôme de degré inférieur ou égal à $2\mathrm{N}+1$ pour un $n \in \mathbb{N}$ donné.

Exercice 24 — Équations différentielles.

1. Montrer le lemme préliminaire suivant : étant donné un réel $\beta > 0$ et deux suites (θ_n) et (α_n) de réels positifs ou nuls vérifiant :

$$\forall n \geqslant 0, \quad \theta_{n+1} \leqslant (1+\beta)\theta_n + \alpha_n \tag{9.76}$$

alors :

$$\forall n \geqslant 0, \quad \theta_n \leqslant \mathrm{e}^{n\beta} + \sum_{i=0}^{n-1} \mathrm{e}^{(n-i-1)\beta}\,\alpha_i \tag{9.77}$$

2. Soit $x_0 \in \mathbb{R}$, $a > 0$ et $f \in \mathcal{C}^0([x_0\,;x_0+a]\times\mathbb{R};\mathbb{R})$ vérifiant la condition de Lipschitz :

$$\exists L > 0 \text{ tel que } \forall x \in [x_0\,;x_0+a],\ \forall y \text{ et } z \in \mathbb{R},\ |f(x,y)-f(x,z)| \leqslant |y-z| \quad (9.78)$$

On considère l'équation différentielle :

$$\begin{cases} y'(x) = f(x,y(x)), \quad x \in [x_0\,;x_0+a] \\ y(x_0) = \alpha \quad \text{donné} \end{cases} \quad (9.79)$$

Ce problème admet une solution $y \in \mathcal{C}^0([x_0\,;x_0+a])$ et une seule. On cherche une approximation numérique de cette solution en n points x_i de l'intervalle $[x_0\,;x_0+a]$ définis par $x_i = x_0 + ih, i = 0,\dots,n$ avec $hn = a$. On calcule, pour $i = 0,\dots,n$, une valeur approchée y_i de $y(x_i)$ au point x_i par le schéma suivant :

$$\begin{cases} y_0 \quad \text{donné} \\ y_{i+1} = y_i + hf\left(x_i + \dfrac{h}{2}, y_i + \dfrac{h}{2}f(x_i,y_i)\right), \quad i = 0,\dots,n-1 \end{cases} \quad (9.80)$$

Lorsque $f \in \mathcal{C}^2([x_0\,;x_0+a]\times\mathbb{R};\mathbb{R})$, montrer que la méthode (9.80) est stable et d'ordre 2.

Exercice 25 — Approximation d'une fonction par une spline cubique. Une *spline cubique* est une fonction de classe $\mathcal{C}^2$ polynomiale par morceau, de degré inférieur ou égal à trois.

Soient $f : [a\,;b] \to \mathbb{R}$, continue sur $[a\,;b]$ et $\{x_i = a + ih;\ i = 0,\dots,n\}$, une subdivision uniforme de l'intervalle $[a\,;b]$ de pas $h = (b-a)/n$. On veut montrer l'existence d'une spline cubique S associée à la subdivision $(x_0,\dots,x_n)$ (c'est-à-dire une fonction $S \in \mathcal{C}^2([a\,;b])$ telle que $S_{|[x_i;x_{i+1}]}$ soit un polynôme de degré inférieur ou égal à trois) qui vérifie :

$$\begin{cases} S(x_i) = f(x_i) \quad i = 0,\dots,n \\ S''(a) = 0 \\ S''(b) = 0 \end{cases} \quad (9.81)$$

On notera $y_i = f(x_i)$ pour $i = 0,\dots,n$ et $P_i(x) = a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3$ pour $i = 0,\dots,n-1$.

1. Montrer que les polynômes P_i vérifient les équations suivantes :

$$P_i(x_i) = y_i \text{ et } P_i(x_{i+1}) = y_{i+1}, \quad i = 0,\dots,n-1 \quad (9.82)$$
$$P_0''(x_0) = P_{n-1}''(x_n) = 0 \quad (9.83)$$
$$P_i'(x_i) = P_{i-1}'(x_i), \quad i = 0,\dots,n-1 \quad (9.84)$$
$$P_i''(x_i) = P_{i-1}''(x_i), \quad i = 0,\dots,n-1 \quad (9.85)$$

Vérifier qu'il y a autant d'équations que d'inconnues du problème.
2. Donner la valeur des coefficients a_i en fonction des y_i.
3. Exprimer les coefficients d_i en fonction des c_i et des y_i.
4. Exprimer les coefficients b_i en fonction des c_i et des y_i.
5. Montrer que $c_0 = 0$ et que $\mathbf{c} = (c_i)_{i=1,\dots,n-1}$ est solution d'un système tridiagonal que l'on explicitera.
6. Montrer que ce système est inversible. Conclure.

Méthode des éléments finis en dimension 1

On complète l'exercice 6 par l'étude de l'erreur d'approximation et par l'étude d'un deuxième problème approché obtenu en calculant l'intégrale du second membre par une formule d'intégration numérique.

A.1 Étude de l'erreur d'approximation

A.1.1 Notations

On note :

$$(u,v)_0 = \int_0^1 uv\,\mathrm{d}x \quad \text{et} \quad |u|_0 = \sqrt{(u,u)_0} \tag{A.1}$$

le produit scalaire de $\mathcal{L}^2(]0\,;1[)$ et la norme associée. On rappelle que $(u',v')_0$ est un produit scalaire sur $\mathcal{H}^1_0(]0\,;1[)$. On le note $((u,v))_0$ et on note $||u||_0$, la norme associée :

$$((u,v))_0 = \int_0^1 u'v'\,\mathrm{d}x \quad \text{et} \quad ||u||_0 = \sqrt{(u',u')_0} \tag{A.2}$$

On note $a(u,v)$, le produit scalaire de $\mathcal{H}^1(]0\,;1[)$ et $|u|_1$, la norme associée :

$$a(u,v) = (u,v)_0 + ((u,v))_0 \quad \text{et} \quad |u|_1 = \sqrt{|u|_0^2 + ||u||_0^2} \tag{A.3}$$

Avec ces notations, le problème (9.12) s'écrit :

$$u \in \mathrm{V} \quad \text{et} \quad \forall v \in \mathrm{V}, \quad a(u,v) = (f,v)_0 \tag{A.4}$$

et le problème approché (9.14) s'écrit :

$$u_h \in \mathrm{V}_h \quad \text{et} \quad \forall v_h \in \mathrm{V}_h, \quad a(u_h,v_h) = (f,v_h)_0 \tag{A.5}$$

Vérifier l'inégalité de Poincaré :

$$\forall v \in \mathrm{V}, \quad |v|_0 \leqslant ||v||_0 \tag{A.6}$$

et en déduire l'équivalence des normes :

$$\forall v \in \mathrm{V}, \quad ||v||_0 \leqslant |v|_1 \leqslant \sqrt{2}||v||_0 \tag{A.7}$$

A.1.2 Majoration de $||u - u_h||_0$

Montrer que :

$$\forall v_h \in \mathrm{V}_h, \quad a(u - u_h, v_h) = 0 \tag{A.8}$$

En déduire que :

$$\forall v_h \in \mathrm{V}_h, \quad |u - u_h|_1^2 = a(u - u_h, u - v_h) \tag{A.9}$$

et montrer alors que :

$$\forall v_h \in \mathrm{V}_h, \quad |u - u_h|_1 \leqslant |u - v_h|_1 \tag{A.10}$$

puis :

$$\forall v_h \in \mathrm{V}_h, \quad ||u - u_h||_0 \leqslant \sqrt{2}||u - v_h||_0 \tag{A.11}$$

On va considérer $v_h = \pi_h u$, interpolé de u dans V_h, défini par :

$$\pi_h u \in \mathrm{V}_h, \quad \pi_h u(x_i) = u(x_i), \quad i = 1, \ldots, n \tag{A.12}$$

On se place sur $\mathrm{I}_i = [x_i\,; x_{i+1}]$. Vérifier que :

$$\forall x \in \mathrm{I}_i, \quad (\pi_h u)'(x) = \frac{u(x_{i+1}) - u(x_i)}{h} \quad \text{où} \quad h = x_{i+1} - x_i \tag{A.13}$$

et montrer que :

$$\forall x \in \mathrm{I}_i, \quad |(\pi_h u)'(x) - u'(x)| \leqslant 2\sqrt{h}\Big(\int_{x_i}^{x_{i+1}} |u''(y)|^2 \,\mathrm{d}y\Big)^{1/2} \tag{A.14}$$

à partir de développements de Taylor en x_i avec reste intégral ; en déduire :

$$\int_{x_i}^{x_{i+1}} |(\pi_h u)'(x) - u'(x)|^2 \,\mathrm{d}x \leqslant 4h^2 \int_{x_i}^{x_{i+1}} |u''(y)|^2 \,\mathrm{d}y \tag{A.15}$$

puis, en sommant sur i, montrer que *l'erreur d'interpolation* vérifie :

$$||\pi_h u - u||_0 \leqslant 2h|u''|_0 \tag{A.16}$$

Pour u solution du problème (A.4), montrer que :

$$|u''|_0 \leqslant 2|f|_0 \tag{A.17}$$

et conclure que l'on obtient la majoration suivante :

$$||u - u_h||_0 \leqslant 4\sqrt{2}h|f|_0 \tag{A.18}$$

A.1.3 Majoration de $|u - u_h|_0$

On rappelle la propriété suivante :

$$\forall v \in \mathcal{L}^2(]0\,;1[), \quad |v|_0 = \sup\left\{\frac{|(v,g)_0|}{|g|_0};\ g \in \mathcal{L}^2(]0\,;1[)\right\} \tag{A.19}$$

On considère l'application T qui à $g \in \mathcal{L}^2(]0\,;1[)$ associe la fonction $\mathrm{T}_g \in \mathrm{V}$ définie par :

$$\forall v \in \mathrm{V}, \quad a(\mathrm{T}_g, v) = (g, v)_0 \tag{A.20}$$

On remarque que T_g existe et est unique d'après le théorème de Riesz ; en particulier, $u = \mathrm{T}_f$.
Montrer que pour $g \in \mathcal{L}^2(]0\,;1[)$ et $v_h \in \mathrm{V}_h$, on a :

$$(u - u_h, g)_0 = a(\mathrm{T}_g - v_h, u - u_h) \tag{A.21}$$

et en déduire :

$$|(u - u_h, g)_0| \leqslant 2||\mathrm{T}_g - v_h||_0\, ||u - u_h||_0 \tag{A.22}$$

En utilisant ce résultat avec $v_h = \pi_h \mathrm{T}_g$, montrer que :

$$|u - u_h|_0 \leqslant 8h||u - u_h||_0 \tag{A.23}$$

et conclure que l'on obtient la majoration suivante :

$$|u - u_h|_0 \leqslant 32\sqrt{2}h^2|f|_0 \tag{A.24}$$

A.1.4 Majoration de $|u - u_h|_\infty$

Pour tout $u \in \mathrm{V}$, montrer que :

$$\forall x \in [0\,;1], \quad u(x)^2 = 2\int_0^x u(y)u'(y)\,\mathrm{d}y \tag{A.25}$$

et en déduire que :

$$|u|_\infty \leqslant \sqrt{2\,|u|_0\,||u||_0} \tag{A.26}$$

où $|u|_\infty = \sup\{|u(x)|;\ x \in [0\,;1]\}$. Appliquer ce résultat à $u - u_h$ pour obtenir :

$$|u - u_h|_\infty \leqslant 16\sqrt{2}h^{3/2}\,|f|_0 \tag{A.27}$$

A.2 Problème approché avec intégration numérique

Dans le problème approché (A.5), on en sait en général pas calculer exactement :

$$(f, v_h)_0 = \int_0^1 f v_h\,\mathrm{d}x \tag{A.28}$$

et on pose alors le problème approché avec intégration numérique :

$$\tilde{u}_h \in \mathrm{V}_h \quad \text{et} \quad \forall v_h \in \mathrm{V}_h, \quad a(\tilde{u}_h, v_h) = (f, v_h)_h \tag{A.29}$$

où $(f, v_h)_h$ est une formule d'intégration numérique de $f v_h$ sur $[0\,;1]$.

A.2.1 Majoration de $||u - \tilde{u}_h||_0$

Montrer que :

$$\forall v_h \in \mathrm{V}_h, \quad |\tilde{u}_h - v_h|_1^2 = a(u - v_h, \tilde{u}_h - v_h) + (f, \tilde{u}_h - v_h)_h - (f, \tilde{u}_h - v_h)_0 \quad \text{(A.30)}$$

En déduire que :

$$\forall v_h \in \mathrm{V}_h, \quad |\tilde{u}_h - v_h|_1 \leqslant |\tilde{u} - v_h|_1 + \mathrm{R}_h \quad \text{(A.31)}$$

où :

$$\mathrm{R}_h = \sup\left\{\frac{|(f, w_h)_h - (f, w_h)_0|}{|w_h|_1};\ w_h \in \mathrm{V}_h\right\} \quad \text{(A.32)}$$

À l'aide d'une inégalité triangulaire, en déduire alors :

$$\forall v_h \in \mathrm{V}_h, \quad |u - \tilde{u}_h|_1 \leqslant 2|u - v_h|_1 + \mathrm{R}_h \quad \text{(A.33)}$$

puis :

$$\forall v_h \in \mathrm{V}_h, \quad ||u - \tilde{u}_h||_0 \leqslant 2\sqrt{2}||u - v_h||_0 + \mathrm{R}_h \quad \text{(A.34)}$$

Appliquer ce résultat à $v_h = \pi_h u$ pour obtenir :

$$||u - \tilde{u}_h||_0 \leqslant 8\sqrt{2}h|f|_0 + \mathrm{R}_h \quad \text{(A.35)}$$

Si on compare au résultat (A.18), on observe que l'on a encore un terme en h auquel s'ajoute R_h qui représente l'erreur due à l'intégration numérique.

A.2.2 Majoration de $|u - \tilde{u}_h|_0$

On reprend ce qui a été fait pour majorer $|u - u_h|_0$ en sous-section A.1.3.

Montrer ici que pour $g \in \mathcal{L}^2(]0\,;1[)$ et $v_h \in \mathrm{V}_h$, on a :

$$(u - \tilde{u}_h, g)_0 = a(\mathrm{T}_g - v_h, u - \tilde{u}_h) + (f, v_h)_0 - (f, v_h)_h \quad \text{(A.36)}$$

et en déduire :

$$|(u - \tilde{u}_h, g)_0| \leqslant 2||\mathrm{T}_g - v_h||_0\,||u - \tilde{u}_h||_0 + \sqrt{2}\mathrm{R}_h||v_h||_0 \quad \text{(A.37)}$$

On considère l'inégalité triangulaire :

$$||v_h||_0 \leqslant 2||\mathrm{T}_g - v_h||_0 + ||\mathrm{T}_g||_0 \quad \text{(A.38)}$$

Montrer que :

$$||\mathrm{T}_g||_0 \leqslant |g|_0 \quad \text{(A.39)}$$

En choisissant $v_h = \pi_h \mathrm{T}_g$, montrer alors que :

$$|u - \tilde{u}_h|_0 \leqslant 8h||u - \tilde{u}_h||_0 + 5\sqrt{2}\mathrm{R}_h \quad \text{(A.40)}$$

Conclure que l'on obtient la majoration suivante :

$$|u - \tilde{u}_h|_0 \leqslant 64\sqrt{2}h^2|f|_0 + (8 + 5\sqrt{2})\mathrm{R}_h \quad \text{(A.41)}$$

En comparant au résultat (A.27), on observe que l'on a encore un terme en h^2 auquel s'ajoute un terme en R_h dû à l'intégration numérique.

A.2.3 Erreur d'intégration numérique par la méthode des trapèzes

Sur chaque intervalle I_i, on écrit la formule des trapèzes :

$$\int_{x_i}^{x_{i+1}} g(x)\,\mathrm{d}x = \frac{h}{2}(g(x_i) + g(x_{i+1})) + \mathrm{E}_i(h) \tag{A.42}$$

où l'erreur commise vaut :

$$\mathrm{E}_i(h) = \int_{x_i}^{x_{i+1}} g(x)\,\mathrm{d}x - \frac{h}{2}(g(x_i) + g(x_{i+1})) \tag{A.43}$$

écrire un développement de Taylor de $\mathrm{E}_i(h)$ à l'ordre 2, en $h = 0$, avec reste intégral (on suppose que $g \in \mathcal{H}^2(]0\,;1[)$). Montrer que :

$$|\mathrm{E}_i(h)| \leqslant \frac{h^2}{2} \int_{x_i}^{x_{i+1}} |g''(x)|\,\mathrm{d}x \tag{A.44}$$

En sommant sur i, on obtient la formule composée des trapèzes :

$$\int_0^1 g(x)\,\mathrm{d}x = h\left(\frac{1}{2}g(0) + \sum_{i=1}^{n-1} g(x_i) + \frac{1}{2}g(1)\right) + \mathrm{E}(h) \tag{A.45}$$

Montrer que l'erreur $\mathrm{E}(h)$ vérifie :

$$|\mathrm{E}(h)| \leqslant \frac{h^2}{2}|g''|_0 \tag{A.46}$$

A.2.4 Majoration de R_h si on utilise la méthode des trapèzes

On suppose que $f \in \mathcal{C}^2([0\,;1])$. Par la formule composée des trapèzes, on a :

$$\forall v_h \in \mathrm{V}_h, \quad (f, v_h)_h = h \sum_{i=1}^{n-1} f(x_i) v_h(x_i) \tag{A.47}$$

avec, d'après ce qui précède au paragraphe A.2.3 :

$$\forall w_h \in \mathrm{V}_h, \quad |(f, w_h)_h - (f, w_h)_0| \leqslant \frac{h^2}{2}|(f\ w_h)''|_0 \tag{A.48}$$

Montrer que :

$$|(f\ w_h)''|_0 \leqslant |f''|_\infty\, |w_h|_0 + 2|f'|_\infty\, ||w_h||_0 \tag{A.49}$$

et en déduire que :

$$\mathrm{R}_h \leqslant \frac{h^2}{2}(|f''|_\infty + 2|f'|_\infty) \tag{A.50}$$

Montrer qu'il existe des constantes C_0, C_1 et C_2, fonctions de la donnée f du problème, telles que :

$$||u - \tilde{u}_h||_0 \leqslant \mathrm{C}_0\, h \tag{A.51}$$

$$|u - \tilde{u}_h|_0 \leqslant \mathrm{C}_1\, h^2 \tag{A.52}$$

$$|u - \tilde{u}_h|_\infty \leqslant \mathrm{C}_2\, h^{3/2} \tag{A.53}$$

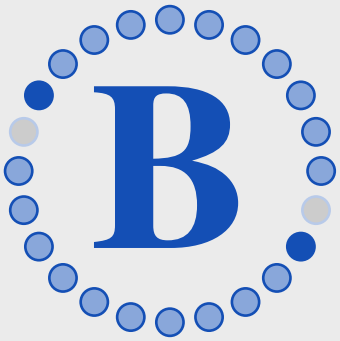
Comparer aux majorations d'erreur obtenues pour le problème approché sas intégration numérique.

A.3 Tests numériques de résolution de problèmes approchés

On choisit plusieurs fonctions $u \in \mathrm{V}$ et on calcule la donnée f correspondante du problème 9.10. On obtient ainsi des exemples de problèmes dont on connaît la solution exacte u :

$$u(x) = x(x-1) \tag{A.54}$$
$$u(x) = (x-1)\sin x \tag{A.55}$$
$$u(x) = \sin(\pi x) \tag{A.56}$$

1. Écrire les algorithmes, programmes et sous-programmes Fortran correspondants pour résoudre le problème approché avec intégration numérique.
2. Résoudre le problème approché sans intégration numérique pour l'exemple (A.54).
3. Avec différentes jeux d'essais, vérifier la convergence des solutions approchées vers les solutions exactes.



Méthode de la puissance itérée pour le calcul de valeurs propres

L'objet de ce travail est l'étude d'une classe de méthodes numériques de calcul de quelques valeurs propres d'une matrice réelle donnée et de vecteurs propres associés.

Ces méthodes entrent dans le cadre des méthodes dites de *la puissance itérée*.

B.1 Itérations simples

▷ **Problème B.1** Soit $\mathbf{A} \in \mathbb{R}^{n,n}$, une matrice réelle d'ordre n supposée diagonalisable dans $\mathbb{C}$. On notera $\lambda_i, i = 1, \ldots, n$ ses valeurs propres dans $\mathbb{C}$ classées dans l'ordre décroissant des modules et $(\mathbf{u}_i)$ une base de vecteurs propres associés. On suppose tout d'abord que les valeurs propres de $\mathbf{A}$ vérifient :

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n| \tag{B.1}$$

On cherche à calculer numériquement la valeur propre de plus grand module λ_1 (appelée valeurs propres principale de $\mathbf{A}$) et un vecteur propre associé, c'est-à-dire un vecteur $\mathbf{x} \neq \mathbf{0}$ tel que $\mathbf{A}\mathbf{x} = \lambda_1 \mathbf{x}$.

B.1.1 Résultats généraux

Montrer que, sous les hypothèses du problème B.1, la valeur propre λ_1 est réelle et que l'on peut trouver un vecteur propre associé qui est réel.

B.1.2 Approximation d'un vecteur propre associé à λ_1

1. Étant donné un vecteur $\mathbf{x}^{(0)}$ non nul de $\mathbb{R}^n$, s'assurer que l'on peut décomposer $\mathbf{x}^{(0)}$ dans $\mathbb{C}$ sur une base de vecteurs propres de $\mathbf{A}$ que l'on notera $u_1, u_2, \ldots, u_n$. Exprimer ce que signifie sur ces coefficients l'hypothèse suivante :

▷ **Hypothèse 1** On suppose que $\mathbf{x}^{(0)}$ n'est pas orthogonal à l'espace propre à gauche associé à la valeur propre λ_1.

2. Soit donc $\mathbf{x}^{(0)}$ un vecteur non nul de $\mathbb{R}^n$. On construit par récurrence à partir de $\mathbf{x}^{(0)}$, une suite de vecteurs réels $\mathbf{x}^{(k)}$, $k \in \mathbb{N}$, par la relation :

$$\mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{x}^{(k)} \tag{B.2}$$

Exprimer $\mathbf{x}^{(k)}$ sur la base (u_i) puis $x_j^{(k)}$ pour $j = 1, \ldots, n$ où $x_j^{(k)}$ désigne la j^{e} composante du vecteur $\mathbf{x}^{(k)}$ et $\mathbf{x}^{(k)}/\lambda_1^k$.

3. En déduire que si la matrice $\mathbf{A}$ vérifie les hypothèses du problème B.1 et si $\mathbf{x}^{(0)}$ vérifie l'hypothèse 1 alors *la suite* $\mathbf{x}^{(k)}/\lambda_1^k$ *converge vers un vecteur propre* (que l'on précisera) associé à λ_1. Montrer que le *facteur de convergence* est de l'ordre de $|\lambda_2|/|\lambda_1|$.

■ **Exemple 1** Soit une matrice $\mathbf{A}_1$ de $\mathbb{R}^{2,2}$ et le vecteur initial $\mathbf{x}^{(0)}$ définis par :

$$\mathbf{A}_1 = \begin{bmatrix} -4 & -6 \\ -6 & -4 \end{bmatrix} \quad \text{et} \quad \mathbf{x}^{(0)} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \tag{B.3}$$

Calculer les premiers itérés $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$ définis par la relation (B.2) ; observer et commenter les résultats.

Faire de même avec la matrice :

$$\mathbf{A}_2 = \begin{bmatrix} -1 & -9 \\ -9 & -1 \end{bmatrix} \tag{B.4}$$

et le même vecteur initial $\mathbf{x}^{(0)}$. Calculer les facteurs de convergence pour $\mathbf{A}_1$ et $\mathbf{A}_2$. Observer la différence de rapidité de convergence entre ces deux cas.

Remarque Les résultats peuvent être différents suivant les moyens de calcul utilisés et en particulier suivant le nombre de chiffres significatifs conservés.

B.1.3 Approximation de la valeur propre λ_1

On considère toujours :

— une matrice $\mathbf{A}$ vérifiant les hypothèses du problème B.1 ;
— un vecteur $\mathbf{x}^{(0)}$ vérifiant l'hypothèse 1 ;
— une suite de vecteurs réels $\mathbf{x}^{(k)}$ ($k \in \mathbb{N}$) définie par la relation (B.2).

Montrer l'existence d'au moins un indice $j \in \{1, \ldots, n\}$ tel que :

$$\lim_{k \to +\infty} \frac{x_j^{(k+1)}}{x_j^{(k)}} = \lambda_1 \tag{B.5}$$

Quel est le facteur de convergence ?

■ **Exemple 2** Observer cette convergence sur les résultats de l'exemple 1.

B.1.4 Cas où l'itéré initial est orthogonal à l'espace propre à gauche associé à λ_1

On suppose toujours que la matrice $\mathbf{A}$ vérifie les hypothèses du problème B.1, avec plus précisément, à la place de (B.16) :

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n| \tag{B.6}$$

On suppose ici que l'hypothèse 1 n'est *pas* vérifiée mais que, par contre, on a :

▷ **Hypothèse 2** On suppose que $\mathbf{x}^{(0)}$ n'est pas orthogonal à l'espace propre à gauche associé à la valeur propre λ_2.

Reprendre les calculs des questions précédentes. Quels résultats de *convergence* obtient-on ? (Convergence vers un vecteur propre, convergence vers une valeur propre).

Remarque En pratique, les calculs sont faits avec des erreurs d'arrondi qui entraînent que certaines quantités nulles théoriquement sont souvent non nulles *numériquement*. Quel résultat numérique peut-on donc espérer pour les limites des suites $\mathbf{x}^{(k)}/\lambda_1^k$ et $x_j^{(k+1)}/x_j^{(k)}$ sous les hypothèses du paragraphe B.1.4 ?

■ **Exemple 3** Soit la matrice :

$$\mathbf{A}_3 = \begin{bmatrix} 18 & -8 & -16 \\ 17 & -7 & -16 \\ -0{,}5 & 0{,}5 & 2 \end{bmatrix} \tag{B.7}$$

Calculer $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ en partant successivement des vecteurs :

$$\mathbf{x}^{(0)} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \quad ; \quad \mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 10^{-5} \\ 1 \end{pmatrix} \quad ; \quad \mathbf{x}^{(0)} = \begin{pmatrix} 10^{-5} \\ -2 \\ 1 \end{pmatrix} \tag{B.8}$$

Qu'observe-t-on ? Noter que l'on a $\mathbf{A}_3 = \mathbf{PDP}^{-1}$ avec :

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad ; \quad \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad ; \quad \mathbf{P}^{-1} = \begin{bmatrix} -0{,}5 & 0{,}5 & 0 \\ 2 & -1 & -2 \\ -0{,}5 & 0{,}5 & 1 \end{bmatrix} \tag{B.9}$$

B.1.5 Amélioration de la méthode

On remarquera que si $|\lambda_1|$ est grand, les termes calculés peuvent vite devenir très grands. De même, si $|\lambda_1|$ est petit, les termes calculés peuvent vite devenir très petits. Les calculs risquent de devenir impossibles du fait de la perte de précision. Pour remédier à cela, on *normalise* les vecteurs calculés de la façon suivante : étant donnée une norme vectorielle $||\cdot||$ définie sur $\mathbb{R}^n$, la suite $(\mathbf{x}^{(k)})$ définie en (B.2) est remplacée par la suite :

$$\begin{cases} \mathbf{q}^{(k)} = \mathbf{A}\mathbf{x}^{(k)} \\ \mathbf{x}^{(k+1)} = \mathbf{q}^{(k)} / ||\mathbf{q}^{(k)}|| \end{cases} \tag{B.10}$$

1. Montrer que, lorsque la suite $(\mathbf{x}^{(k)})$ est bien définie, on a :

$$||\mathbf{x}^{(k)}|| = 1 \quad ; \quad \mathbf{x}^{(k)} = \frac{\mathbf{A}^k \mathbf{x}^{(0)}}{||\mathbf{A}^k \mathbf{x}^{(0)}||} \tag{B.11}$$

2. Montrer alors, sous les hypothèses sur $\mathbf{A}$ du problème B.1 et si l'hypothèse 1 est vérifiée, que :
   — la suite :

$$\left(\frac{\lambda_1}{|\lambda_1|}\right)^k \mathbf{x}^{(k)} \tag{B.12}$$

   converge vers un vecteur propre associé à λ_1 ;
   — il existe au moins un indice $j \in \{1, \ldots, n\}$ tel que :

$$\lim_{k \to +\infty} \frac{(\mathbf{A}\mathbf{x}^{(k)})_j}{\mathbf{x}_j^{(k)}} = \lambda_1 \tag{B.13}$$

   Interpréter suivant le signe de λ_1.
3. Dans quel cas la suite $(\mathbf{x}^{(k)})$ n'est-elle pas définie ?

■ **Exemple 4** Reprendre les calculs faits pour la matrice $\mathbf{A}_1$ en normalisant les vecteurs $\mathbf{x}^{(k)}$ à chaque étape.

B.1.6 Calcul d'autres éléments propres : méthode de déflation

On suppose que la matrice $\mathbf{A}$ vérifie les hypothèses du problème B.1. Soit $\mathbf{v}_1$ un vecteur propre à gauche de $\mathbf{A}$ associé à λ_1.

1. Montrer que la matrice $\mathbf{B} \in \mathbb{R}^{n,n}$ définie par :

$$\mathbf{B} = \mathbf{A} - \frac{\lambda_1}{\mathbf{v}_1^\top \mathbf{u}_1} \mathbf{u}_1 \mathbf{v}_1^\top \tag{B.14}$$

où $\mathbf{v}_1^\top$ désigne le vecteur transposé de $\mathbf{v}_1$, admet pour valeurs propres $0, \lambda_2, \ldots, \lambda_n$ associées, respectivement, aux vecteurs propres $\mathbf{u}_1, \ldots, \mathbf{u}_n$.

■ **Exemple 5** Calculer la matrice déflatée $\mathbf{B}$ pour la matrice $\mathbf{A}_3$.

2. En déduire une méthode de calcul de toutes les valeurs propres d'une matrice réelle dont les valeurs propres sont telles que :

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| \tag{B.15}$$

Préciser comment on peut calculer numériquement le vecteur $\mathbf{v}_1$.

■ **Exemple 6** Calculer un vecteur $\mathbf{v}_1$ associé à la matrice $\mathbf{A}_3$.

3. On suppose que la matrice $\mathbf{A}$ est symétrique : quelle simplification cette hypothèse apporte-t-elle à l'étude de 2., notamment pour le calcul numérique de $\mathbf{v}_1$?

■ **Exemple 7** Calculer la matrice déflatée $\mathbf{B}$ pour la matrice $\mathbf{A}_1$.

B.2 Méthode d'accélération de convergence

▷ **Problème B.2** Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice réelle d'ordre n supposée symétrique. On notera λ_i, $i = 1, \ldots, n$ ses valeurs propres dans $\mathbb{C}$ classées dans l'ordre décroissant des modules et vérifiant :

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n| \tag{B.16}$$

On cherche à calculer numériquement la valeur propre λ_1.

Soit $\mathbf{x}^{(0)}$ un vecteur non nul de $\mathbb{R}^n$. On construit par récurrence, à partir de $\mathbf{x}^{(0)}$, une suite de réels R_k par :

$$\begin{cases} \mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{x}^{(k)} \\ \mathrm{R}_k = \dfrac{\mathbf{x}^{(k)\top}\mathbf{x}^{(k+1)}}{||\mathbf{x}^{(k)}||_2^2} \end{cases} \tag{B.17}$$

Exprimer $\mathbf{x}^{(k)}$ et $\mathbf{x}^{(k+1)}$ sur une base orthonormée $(\mathbf{u}_i)$ de $\mathbb{C}^n$. Montrer que si $\mathbf{x}^{(0)}$ n'est pas orthogonal à l'espace propre associé à λ_1 alors :

$$\lim_{k \to +\infty} \mathrm{R}_k = \lambda_1 \tag{B.18}$$

avec un facteur de convergence de l'ordre de $(|\lambda_2|/|\lambda_1|)^2$.

■ **Exemple 8** Calculer $\mathrm{R}_0, \mathrm{R}_1, \ldots$ pour la matrice $\mathbf{A}_1$. Observer la rapidité de convergence.

B.3 Méthode de la puissance itérée inverse

▷ **Problème B.3** Soit $\mathbf{A} \in \mathbb{R}^{n,n}$ une matrice réelle d'ordre n supposée inversible et diagonalisable dans $\mathbb{C}$. On notera toujours λ_i, $i = 1, \ldots, n$ ses valeurs propres dans $\mathbb{C}$ classées dans l'ordre décroissant des modules et on cherche à calculer numériquement la valeur propre λ_n de plus petit module.

L'idée générale est la suivante : appliquer la méthode de la puissance itérée à $\mathbf{A}^{-1}$ (la plus grande valeur propre de $\mathbf{A}^{-1}$ est la plus petite valeur propre de $\mathbf{A}$) mais, au lieu de calculer les itérés à l'aide de (B.2) avec $\mathbf{A}^{-1}$ (qui nécessite le calcul de $\mathbf{A}^{-1}$), on résout à chaque étape le système linéaire :

$$\mathbf{A}\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} \tag{B.19}$$

En particulier, lorsque $\mathbf{A}$ admet une décomposition en un produit $\mathbf{LU}$ sans permutations, on calcule une fois $\mathbf{L}$ et $\mathbf{U}$, puis on détermine la suite $\mathbf{x}^{(k)}$ en résolvant successivement, à chaque étape k, deux systèmes linéaires : pour $\mathbf{x}^{(0)}$ donné dans $\mathbb{R}$,

$$\begin{cases} \mathbf{L}\mathbf{y} = \mathbf{x}^{(k)} \\ \mathbf{U}\mathbf{x}^{(k+1)} = \mathbf{y} \end{cases} \tag{B.20}$$

(Ce procédé est plus stable et nécessite moins d'opérations.)

1. Écrire des conditions suffisantes pour que la méthode itérative (B.20) soit convergente et donne des approximations de λ_n et d'un vecteur propre associé.

■ **Exemple 9** Calculer par ce procédé la valeur propre λ_2 pour la matrice $\mathbf{A}_1$. Indication : on a $\mathbf{A}_1 = \mathbf{LU}$ avec :

$$\mathbf{L} = \begin{bmatrix} 1 & 0 \\ 1{,}5 & 1 \end{bmatrix} \quad ; \quad \mathbf{U} = \begin{bmatrix} -4 & -6 \\ 0 & 5 \end{bmatrix} \tag{B.21}$$

2. Soit $\mathbf{A}$ une matrice réelle donnée d'ordre n supposée inversible, diagonalisable dans $\mathbf{C}$ et dont on connaît une approximation d'une valeur propre λ_i supposée isolée, c'est-à-dire $|\lambda_{i-1}| > |\lambda_i| > |\lambda_{i+1}|$. Comment peut-on utiliser le procédé itératif (B.20) pour calculer un vecteur propre associé à λ_i ? On regardera avec attention l'inversibilité des systèmes linéaires.

■ **Exemple 10** À partir de l'approximation $\mu_2 = 2{,}1$ de la valeur propre λ_2 de la matrice $\mathbf{A}_1$, calculer plus précisément λ_2 et un vecteur propre associé. Indication : on a $\mathbf{A}_3 - \mu_2\mathbf{I} = \mathbf{LU}$ avec :

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1{,}0692 & 1 & 0 \\ -0{,}0314 & -0{,}4545 & -1 \end{bmatrix} \quad ; \quad \mathbf{U} = \begin{bmatrix} 15{,}9 & -8 & -16 \\ 0 & -0{,}5465 & 1{,}1069 \\ 0 & 0 & -0{,}1 \end{bmatrix} \tag{B.22}$$

On suppose maintenant que les hypothèses précédentes ne sont plus vérifiées.

■ **Exemple 11** On a $\lambda_1 = \lambda_2$ et $|\lambda_2| > |\lambda_3|$:

$$\mathbf{A}_5 = \begin{bmatrix} 18 & 8 & -16 \\ 8 & 2 & -16 \\ 4 & -4 & 2 \end{bmatrix} \tag{B.23}$$

■ **Exemple 12** On a $\lambda_1 = -\lambda_2$ et $|\lambda_2| > |\lambda_3|$:

$$\mathbf{A}_5 = \begin{bmatrix} 18 & 8 & -16 \\ 28 & -18 & -16 \\ -6 & 6 & 2 \end{bmatrix} \tag{B.24}$$

■ **Exemple 13** On a $|\lambda_1| = |\lambda_2|$ avec λ_1 et $\lambda_2 \in \mathbb{C}$ et $|\lambda_2| > |\lambda_3|$:

$$\mathbf{A}_6 = \begin{bmatrix} 5{,}5 & -1{,}5 & -6 \\ -10{,}5 & 8{,}5 & 6 \\ 7{,}5 & -4{,}5 & -5 \end{bmatrix} \tag{B.25}$$

Prérequis d'analyse numérique

C.1 Analyse matricielle

Exercice 26 Soient $a \in \mathbb{R}$ et n, m, p trois entiers tels que $1 \leqslant m, p \leqslant n$. Soit $\mathbf{M} = (m_{ij})$ la matrice carrée d'ordre n définie par :

$$m_{ij} = \delta_{ij} + a\delta_{mi}\delta_{pj}, \quad i, j = 1, \ldots, n \tag{C.1}$$

La matrice $\mathbf{M}$ est-elle inversible ? Si oui, calculer son inverse.

Exercice 27 On considère une matrice de Jordan d'ordre n :

$$\mathbf{J}_{\lambda,n} = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \lambda \end{bmatrix} \tag{C.2}$$

Calculer les puissances $\mathbf{J}_{\lambda,n}^m$ de $\mathbf{J}_{\lambda,n}$ pour un entier $m \geqslant 1$. On pourra décomposer $\mathbf{J}_{\lambda,n}$ en $\mathbf{J}_{\lambda,n} = \lambda\mathbf{I} + \mathbf{N}$ où $\mathbf{I}$ est la matrice identité de $\mathbb{R}^{n,n}$ et $\mathbf{N}$, une matrice nilpotente.

Exercice 28 Soit $\mathbf{A}$ une matrice réelle symétrique définie positive. Montrer, en utilisant une décomposition matricielle de $\mathbf{A}$ suivant une matrice diagonale, l'existence d'une matrice $\mathbf{B}$ symétrique définie positive telle que $\mathbf{B}^2 = \mathbf{A}$.

C.2 Algèbre linéaire

Exercice 29 Soit $\mathcal{P}_n$ l'espace vectoriel des polynômes à une variable à coefficients réels et de degré inférieur ou égal à n. Écrire la matrice de l'application linéaire f de $\mathcal{P}_3$ dans $\mathcal{P}_4$ définie par :

$$\forall \mathrm{P} \in \mathcal{P}_3, \quad f(\mathrm{P})(x) = x^2\mathrm{P}'(x) - 2\mathrm{P}(x) \tag{C.3}$$

relativement aux bases canoniques de $\mathcal{P}_3$ et $\mathcal{P}_4$.

Exercice 30 Montrer que la famille de fonctions (h_α) de $\mathbb{R}$ dans $\mathbb{R}$ où $\alpha \in \mathbb{R}$ définies par :

$$h_\alpha : x \in \mathbb{R} \to h_\alpha(x) = |x - \alpha| \tag{C.4}$$

est libre dans $\mathcal{F}(\mathbb{R}, \mathbb{R})$.

Exercice 31 Soit T la transformation de $\mathbb{R}^3$ dans $\mathbb{R}^4$ définie par :

$$\mathrm{T} : \mathbf{x} = (x_j)_{j=1,\ldots,3} \rightarrow \mathrm{T}(\mathbf{x}) = (y_i)_{i=1,\ldots,4} \tag{C.5}$$

avec :

$$\begin{aligned} y_1 &= 5x_1 + 2x_2 - x_3 \\ y_2 &= -8x_1 - 3x_2 + 2x_3 \\ y_3 &= -x_1 - 2x_2 - 3x_3 \\ y_4 &= 3x_1 - x_2 - 5x_3 \end{aligned} \tag{C.6}$$

1. Montrer que T est linéaire.
2. Montrer que l'image du plan P de $\mathbb{R}^3$ d'équation $x_1 + x_2 + x_3 = 0$ est une droite que l'on précisera.

Exercice 32 Soit $\mathbf{u}$ un vecteur (colonne) de $\mathbb{R}$ et de norme euclidienne unitaire, $||\mathbf{u}||_2 = 1$:
1. Calculer les produits $\mathbf{u}\,\mathbf{u}^\top$ et $\mathbf{u}^\top\,\mathbf{u}$ où $\mathbf{u}^\top$ désigne le transposé de $\mathbf{u}$.
2. Montrer que la matrice $\mathbf{H} = \mathbf{I} - 2\mathbf{u}\,\mathbf{u}^\top$, où $\mathbf{I}$ désigne la matrice identité de $\mathbb{R}^{n,n}$, est symétrique et orthogonale.
3. Montrer que la matrice $\mathbf{H}$ définie en 2. admet -1 comme valeur propre simple associée au vecteur propre $\mathbf{u}$ et $+1$ comme valeur propre d'ordre $n-1$ dont le sous-espace propre associé est l'hyperplan orthogonal à $\mathbf{u}$.

C.3 Valeurs propres

Exercice 33 Calculer les valeurs propres de la matrice $\mathbf{A} \in \mathbb{R}^{n,n}$ avec $n \in \mathbb{N}$ et $n > 1$ suivante :

$$\mathbf{A} = (a_{ij})_{i,j=1,\ldots,n} \quad \text{avec} \quad a_{ij} = 1 - \delta_{ij} \tag{C.7}$$

La matrice est-elle inversible ?

Exercice 34 Montrer que les valeurs propres d'une matrice hermitienne (ou d'une matrice réelle symétrique) sont réelles.

Exercice 35 — Théorème de Gershgörin-Hadamard. Soit $\mathbf{A} \in \mathbb{C}^{n,n}$ une matrice carrée.
1. Montrer que toutes les valeurs propres de $\mathbf{A}$ appartiennent à la réunion $\cup_{i=1}^{n} \mathrm{K}_i$ des disques K_i définis par :

$$\mathrm{K}_i = \left\{ z \in \mathbb{C};\ |z - a_i| \leqslant \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\}, \quad i = 1, \ldots, n \tag{C.8}$$

2. Déduire de la question précédente que la matrice $\mathbf{A}_1$ suivante est inversible :

$$\mathbf{A}_1 = \begin{bmatrix} 4 & -1 & 2i & 0 \\ 0 & 2+2i & 1 & 1 \\ 1+i & 0 & -3 & 1 \\ 0 & 2i & -1 & 1+3i \end{bmatrix} \tag{C.9}$$

et que la matrice $\mathbf{A}_2$ suivante est symétrique définie positive :

$$\mathbf{A}_2 = \begin{bmatrix} 4 & -1 & -1 & -1 \\ -1 & 4 & -1 & -1 \\ -1 & -1 & 4 & -1 \\ -1 & -1 & -1 & 4 \end{bmatrix} \tag{C.10}$$

C.4 Résolution numérique de systèmes linéaires

On rappelle la méthode de Gauss sans permutation en s'inspirant de la section 2.1.2. C'est une méthode d'élimination permettant de résoudre un système linéaire $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{n,n}$ et $\mathbf{b} \in \mathbb{R}^n$, lorsque $\mathbf{A}$ est inversible.

On pose $\mathbf{A}^{(1)} \equiv \mathbf{A}$ et $\mathbf{b}^{(1)} \equiv \mathbf{b}$. On élimine successivement l'inconnue x_1 dans les $n-1$ dernières équations, puis, à l'étape $p \in \{1, \dots, n-1\}$, l'inconnue x_p dans les $n-p$ dernières équations par les relations de récurrence suivantes : si $a_{pp}^{(p)} \neq 0$, pour $i = p+1, \dots, n$ on multiplie la ligne p par :

$$m_{ip} = a_{ip}^{(p)} / a_{pp}^{(p)} \tag{C.11}$$

et on soustrait la ligne obtenue à la ligne i. On a pour $i = p+1, \dots, n$:

$$a_{ij}^{(p+1)} = a_{ij}^{(p)} - m_{ip} a_{pj}^{(p)}, \quad j = p, \dots, n \tag{C.12}$$

et :

$$b^{(p+1)} = b_i^{(p)} - m_{ip} b_p^{(p)} \tag{C.13}$$

Il est en général important d'envisager une stratégie du pivot en choisissant le pivot $a_{pp}^{(p)}$ le plus grand possible.

Exercice 36 — Étude d'un système linéaire tridiagonal. Soient n coefficients réels $\alpha_1, \dots, \alpha_n$ et $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$.

1. Écrire, en les justifiant, des conditions suffisantes sur les coefficients (α_i) pour que la matrice $\mathbf{A}_\alpha$, d'ordre n, suivante soit inversible :

$$\mathbf{J}_{\lambda,n} = \begin{bmatrix} \alpha_1 & 1 & 0 & \cdots & 0 \\ 1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & \alpha_n \end{bmatrix} \tag{C.14}$$

2. Pour la résolution d'un système linéaire $\mathbf{A}_\alpha \mathbf{x} = \mathbf{b}$, montrer que lorsque l'algorithme de Gauss peut s'appliquer sans permutations ni de lignes, ni de colonnes, cela revient à calculer n coefficients $(\beta_i)_{i=1,\dots,n}$ de la matrice triangulaire, dont on donnera la formule de récurrence. On démontrera que cet algorithme est possible, par exemple, lorsque les coefficients (α_i) vérifient $\alpha_i > 2$ pour $i = 1, \dots, n$.
3. Donner le nombre total d'opérations pour la résolution du système linéaire.

Exercice 37 — Méthode de Gauss par blocs.

1. Écrire la forme générale du produit matriciel par blocs de deux matrices $\mathbf{A}$ et $\mathbf{B}$ lorsque $\mathbf{A}$ est décomposée en N × P blocs et le produit $\mathbf{C} = \mathbf{AB}$ en N × M blocs.
2. Écrire la méthode de Gauss par blocs pour la résolution d'un système linéaire :

$$\mathbf{Ax} = \mathbf{b}, \quad \mathbf{A} \in \mathbb{R}^{n,n}, \quad \mathbf{b} \in \mathbb{R}^n \tag{C.15}$$

en utilisant une décomposition de $\mathbf{A}$ en quatre blocs. On précisera les dimensions des blocs et des conditions suffisantes sur ces blocs pour que la méthode soit applicable.

Exercice 38 — Étude matricielle de la méthode de Gauss.

1. Montrer que, lorsque la méthode de Gauss peut être effectuée sans permutations, ni de lignes ni de colonnes, sur une matrice $\mathbf{A} \in \mathbb{R}^{n,n}$, cet algorithme revient :
   — à l'étape p : à factoriser la matrice $\mathbf{A}_p$ en un produit :

$$\mathbf{A}_p = \mathbf{L}_p \mathbf{A}_{p+1} \tag{C.16}$$

   où $\mathbf{L}_p$ est une matrice triangulaire inférieure ayant des 1 sur la diagonale ;
   — à l'étape $n-1$: à factoriser la matrice $\mathbf{A}$ en un produit :

$$\mathbf{A} = \mathbf{L}\mathbf{U} \tag{C.17}$$

   où $\mathbf{L}$ est une matrice triangulaire inférieure ayant des 1 sur la diagonale et $\mathbf{U}$, une matrice triangulaire supérieure.
2. Étudier la décomposition matricielle de la méthode lorsque l'on effectue des permutations de lignes ou de colonnes.

Exercice 39 — Étude matricielle du procédé d'orthogonalisation de Gram-Schmidt. On note $||\cdot||_2$ la norme euclidienne sur $\mathbb{R}^n$ et $(\cdot,\cdot)_2$, le produit scalaire associé. Soient m vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_m$ de $\mathbb{R}^n$ supposés linéairement indépendants. À partir de ces m vecteurs, on construit m vecteurs $\mathbf{q}_i$, $i = 1, \ldots, m$, orthonormés pour le produit scalaire $(\cdot,\cdot)_2$, en utilisant le procédé d'orthogonalisation de Gram-Schmidt. On note $\mathbf{A} \in \mathbb{R}^{n,m}$, la matrice dont les colonnes sont les vecteurs $(\mathbf{a}_i)$ et $\mathbf{Q}$, la matrice dont les colonnes sont les vecteurs $(\mathbf{q}_i)$.

Rappel du procédé

— à l'étape $p = 1$, on considère le vecteur :

$$\mathbf{b}_1 = \mathbf{a}_1 \tag{C.18}$$

puis on définit $r_1 = ||\mathbf{b}_1||_2$ et on choisit :

$$\mathbf{q}_1 = \mathbf{b}_1 / r_1 \tag{C.19}$$

— à l'étape $p+1$ ($p \in \{1, \ldots, m-1\}$), on suppose connus p vecteurs $\mathbf{q}_i$ ($i = 1, \ldots, p$) vérifiant :

$$\begin{cases} ||\mathbf{q}_i||_2 = 1 & \text{pour } i = 1, \ldots, p \\ (\mathbf{q}_i, \mathbf{q}_j)_2 = 0 & \text{pour } i > j, i, j = 1, \ldots, p \end{cases} \tag{C.20}$$

On cherche alors un vecteur $\mathbf{b}_{p+1}$ orthogonal aux $\mathbf{q}_i$ ($i = 1, \ldots, p$) de la forme :

$$\mathbf{b}_{p+1} = \mathbf{a}_{p+1} + \alpha_1^p \mathbf{q}_1 + \ldots + \alpha_p^p \mathbf{q}_p \tag{C.21}$$

où les α_i^p sont des coefficients appropriés, puis on définit le vecteur :

$$\mathbf{q}_{p+1} = \mathbf{b}_{p+1} / r_{p+1} \quad \text{où} \quad r_{p+1} = ||\mathbf{b}_{p+1}||_2 \tag{C.22}$$

Question préliminaire Montrer que les conditions d'orthogonalité :

$$(\mathbf{b}_{p+1}, \mathbf{q}_j)_2 = 0, \quad j = 1, \ldots, p \tag{C.23}$$

impliquent :

$$\alpha_j^p = -(\mathbf{a}_{p+1}, \mathbf{q}_j)_2 = 0, \quad j = 1, \ldots, p \tag{C.24}$$

1. Montrer que :

$$\mathbf{A} = \mathbf{QU} \tag{C.25}$$

où $\mathbf{U} \in \mathbb{R}^{m,m}$ est une matrice triangulaire supérieure inversible. On pourra réécrire l'équation (C.18) sous la forme :

$$\mathbf{a}_1 = \beta_1 \mathbf{q}_1 \tag{C.26}$$

et l'équation (C.21) sous la forme :

$$\mathbf{a}_{p+1} = -(\alpha_1^p \mathbf{q}_1 + \ldots + \alpha_p^p \mathbf{q}_p) + \beta_{p+1} \mathbf{q}_{p+1} \tag{C.27}$$

où β_1 et β_{p+1} sont des coefficients à déterminer en considérant la première colonne de $\mathbf{U}$ puis les suivantes.
2. Montrer que $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$ où $\mathbf{I}$ est la matrice identité et $\mathbf{Q}^\top$, la matrice transposée de $\mathbf{Q}$.

▷ **Définition C.1 — Matrice orthogonale.** Une matrice $\mathbf{Q} \in \mathbb{C}^{n,m}$ telle que $\bar{\mathbf{Q}}^\top \mathbf{Q} = \mathbf{I}$ est dite *orthogonale*.

3. Trouver une matrice orthogonale $\mathbf{Q} \in \mathbb{R}^{n,m}$ telle que $\mathbf{Q}\mathbf{Q}^\top \neq \mathbf{I}$.
4. Application à la résolution d'un système linéaire rectangulaire de la forme :

$$\mathbf{Ax} = \mathbf{b} \tag{C.28}$$

lorsque $\mathbf{A} \in \mathbb{R}^{n,m}$, $\mathbf{b} \in \mathbb{R}^n$ avec $\text{rang}(\mathbf{A}) = m$ et $m \leqslant n$.
   (a) Écrire une condition nécessaire et suffisante liant le vecteur $\mathbf{b}$ et les colonnes de $\mathbf{A}$ pour qu'il existe une solution au système. Si cette condition est vérifiée, peut-il y avoir plusieurs solutions ?
   (b) On souhaite résoudre (C.28) en utilisant une décomposition de $\mathbf{A}$ en $\mathbf{QU}$ comme dans la question 1. Supposer tout d'abord l'existence d'une solution du système (C.28) et en déduire alors $\mathbf{x}$ en fonction de $\mathbf{b}$, $\mathbf{Q}$ et $\mathbf{U}$. En déduire une condition nécessaire et suffisante, liant $\mathbf{b}$ et $\mathbf{Q}$, d'existence d'une solution (condition équivalente à celle trouvée en (a)).

Exercice 40 Soit $\mathbf{A} \in \mathbb{R}^{n+1,n+1}$ une matrice carrée, d'ordre $n+1$ de la forme suivante :

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \ldots & 0 & c_1 \\ 0 & 1 & & 0 & c_2 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & c_n \\ b_1 & b_2 & \ldots & b_n & a \end{bmatrix} \tag{C.29}$$

où $a, b_i, i = 1, \ldots, n$ et $c_i, i = 1, \ldots, n$ sont des coefficients réels.

Écrire, lorsque cela est possible, la méthode Gauss sans permutations pour la résolution d'un système linéaire $\mathbf{Ax} = \mathbf{y}$ où $\mathbf{y} \in \mathbb{R}^{n+1}$ est donné. Sous quelle(s) condition(s) sur les coefficients de $\mathbf{A}$ la méthode est-elle applicable ?

Exercice 41 Soient $\mathbf{A}$ une matrice de $\mathbb{R}^{m,n}$ où $n \leqslant m$ et $\mathbf{y}$, un vecteur de $\mathbb{R}^m$. Un vecteur $\mathbf{x} \in \mathbb{R}^n$ est appelée *pseudo-solution* de l'équation :

$$\mathbf{Ax} = \mathbf{y} \tag{C.30}$$

si $\forall \mathbf{z} \in \mathbb{R}^n$, $||\mathbf{Ax} - \mathbf{y}|| \leqslant ||\mathbf{Az} - \mathbf{y}||$ où $||\cdot||$ désigne la norme euclidienne de $\mathbb{R}^m$.

1. Montrer que, si le système linéaire rectangulaire (C.30) admet une solution, alors toute pseudo-solution de (C.30) est une solution de cette équation.
2. Montrer que, pour tout réel λ et tous vecteurs $\mathbf{x}$ et $\mathbf{z} \in \mathbb{R}^n$:

$$||\mathbf{A}(\mathbf{x} + \lambda\mathbf{z}) - \mathbf{y}||^2 = ||\mathbf{A}\mathbf{x} - \mathbf{y}||^2 + 2\lambda(\mathbf{A}\mathbf{z}, \mathbf{A}\mathbf{x} - \mathbf{y}) + \lambda^2||\mathbf{A}\mathbf{z}||^2 \tag{C.31}$$

   En déduire que $\mathbf{x}$ est une pseudo-solution de (C.30) si et seulement si c'est une solution du système linéaire :

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{y} \tag{C.32}$$

3. On suppose maintenant que le rang de $\mathbf{A}$ est égal à n.
   (a) Montrer que la matrice $\mathbf{A}^\top \mathbf{A}$ est une matrice symétrique définie positive. Que peut-on dire de ses valeurs propres ?
   (b) En déduire :
      i. que l'équation (C.32) admet une solution $\mathbf{x}$ et une seule,
      ii. que cette solution $\mathbf{x}$ peut s'écrire sous la forme $\mathbf{x} = \mathbf{B}\mathbf{y}$ où $\mathbf{B}$ est une matrice à préciser et vérifiant $\mathbf{B}\mathbf{A} = \mathbf{I}$, $\mathbf{I}$ étant la matrice identité,
      iii. que la matrice $\mathbf{D} = \mathbf{A}\mathbf{B}$ est symétrique. Donner un exemple de matrice $\mathbf{A}$ pour laquelle $\mathbf{D} = \mathbf{I}$.
   (c) Peut-on utiliser la méthode de Gauss pour la résolution numérique du système (C.32) ? Justifier la réponse.
4. Application : soit P une parabole réelle inconnue d'équation :

$$y(x) = ax^2 + bx + c \tag{C.33}$$

   Cinq mesures expérimentales donnent les résultats suivants :

$$\begin{cases} x_1 = -1 & y_1 = -2 \\ x_2 = 0 & y_2 = 1 \\ x_3 = 1 & y_3 = 1 \\ x_4 = 2 & y_4 = -1{,}5 \\ x_5 = -1 & y_5 = -4 \end{cases} \tag{C.34}$$

   (a) Existe-t-il une parabole P passant par les cinq points $(x_i, y_i)_{i=0,\dots,5}$ donnés en (C.34) ?
   (b) On cherche des constantes a, b et c telles que :

$$\sum_{i=1}^{5} |y_i - y(x_i)|^2 \tag{C.35}$$

   soit minimal.
      i. En utilisant les questions 2. et 3., montrer que ce problème admet au moins une solution. Déterminer une solution $(a, b, c) \in \mathbb{R}^3$ par la même méthode.
      ii. Y a-t-il unicité de la solution ? Justifier la réponse.

C.5 Analyse

Exercice 42 — Polynômes d'interpolation de Lagrange. Soit une fonction $f \in \mathcal{C}^{n+1}([a\,;b])$ et $(x_i)_{i=0,\dots,n}$, $n+1$ points distincts de l'intervalle $[a\,;b]$. On note p_n le polynôme de degré

inférieur ou égal à n d'interpolation de f aux points (x_i). Montrer que pour tout $x \in [a\,;b]$, il existe un point ξ_x appartenant au plus petit intervalle contenant x et les points (x_i) tel que :

$$f(x) - p_n(x) = \frac{1}{(n+1)!}\Pi(x) f^{(n+1)}(\xi_x) \quad \text{où} \quad \Pi(x) = \prod_{i=1}^{n}(x - x_i) \tag{C.36}$$

On étudiera, pour montrer ce résultat, les zéros de la fonction $\mathrm{F} : [a\,;b] \to \mathbb{R}$ définie par :

$$t \to \mathrm{F}(t) = f(t) - p_n(t) - \frac{f(x) - p_n(x)}{\Pi(x)}\Pi(t) \tag{C.37}$$

ainsi que les zéros de ses dérivées.